

Run the cell below to generate the road map (do not modify it)

In [14]:

```
%%javascript
var kernel = IPython.notebook.kernel; var thename =
window.document.getElementById("notebook_name").innerHTML; var command = "THE_NOTEBOOK = " + "'" + thename + "'"; kernel.execute(command); command = "os.environ['THE_NOTEBOOK'] = THE_NOTEBOOK"; kernel.execute(command); var cell = IPython.notebook.get_cell(2); cell.execute(); IPython.notebook.get_cell(3).focus_cell(); var x = $('code_cell'); $(x[1]).children('input').hide();
```

In [15]:

```
outputdir = "/tmp/tools/"
!mkdir -p $outputdir
!wget "https://www.dropbox.com/s/4g0pigmro4volb4/menutemplate?dl=0" -O /tmp/tools/menutemplate >> /tmp/toollog 2>&1
!wget "https://www.dropbox.com/s/3fltpzhsja8td7/construct_menu.py?dl=0" -O /tmp/tools/construct_menu.py >> /tmp/toollog 2>&1
!python /tmp/tools/construct_menu.py "{THE_NOTEBOOK}.ipynb" {outputdir}
from IPython.core.display import HTML
output_file_name = outputdir + THE_NOTEBOOK.replace(" ", "").replace("[", "").replace("]", "") + ".ipynb.html"
with open(output_file_name) as fp:
    html = fp.read()
HTML(html)
```

Out [15]:

Building a music recommender system

As its name implies, a recommender system is a tool that helps predicting what a user may or may not like among a list of given items. In some sense, you can view this as an alternative to content search, as recommendation engines help users discover products or content that they may not come across otherwise. For example, Facebook suggests friends and pages to users. Youtube recommends videos which users may be interested in. Amazon suggests the products which users may need... Recommendation engines engage users to services, can be seen as a revenue optimization process, and in general help maintaining interest in a service.

In this notebook, we study how to build a simple recommender system: we focus on music recommendations, and we use a simple algorithm to predict which items users might like, that is called ALS, alternating least squares.

Goals

In this lecture, we expect students to:

- Revisit (or learn) recommender algorithms

- Understand the idea of Matrix Factorization and the ALS algorithm (serial and parallel versions)
- Build a simple model for a real usecase : music recommender system
- Understand how to validate the results

Steps

In particular, we guide students through the following steps, which constitute a good basis for the end-to-end development of a recommender system:

- Inspect the data using Spark SQL, and build some basic, but very valuable knowledge about the information we have at hand
- Formally define what is a sensible algorithm to achieve our goal: given the "history" of user taste for music, recommend new music to discover. Essentially, we want to build a statistical model of user preferences such that we can use it to "predict" which additional music the user could like
- With our formal definition at hand, we will learn different ways to implement such an algorithm. Our goal here is to illustrate what are the difficulties to overcome when implementing a (parallel) algorithm
- Finally, we will focus on an existing implementation, available in the Apache Spark MLlib, which we will use out of the box to build a reliable statistical model

Now, you may think at this point we will be done!

Well, you'd better think twice: one important topic we will cover in all our Notebooks is **how to validate the results we obtain**, and **how to choose good parameters to train models** especially when using an "opaque" library for doing the job. As a consequence, we will focus on the statistical validation of our recommender system.

1. Data

Understanding data is one of the most important part when designing any machine learning algorithm. In this notebook, we will use a data set published by Audioscrobbler - a music recommendation system for last.fm. Audioscrobbler is also one of the first internet streaming radio sites, founded in 2002. It provided an open API for "scrobbling", or recording listeners' plays of artists' songs. last.fm used this information to build a powerful music recommender engine.

1.1. Data schema

Unlike a rating dataset which contains information about users' preference for products (one star, 3 stars, and so on), the datasets from Audioscrobbler only has information about events: specifically, it keeps track of how many times a user played songs of a given artist and the names of artists. That means it carries less information than a rating: in the literature, this is called explicit vs. implicit ratings.

The data we use in this Notebook is available in 3 files:

- **user_artist_data.txt**: It contains about 141,000 unique users, and 1.6 million unique artists. About 24.2 million users' plays of artists' are recorded, along with their count. It has 3 columns separated by spaces:
-

UserID	ArtistID	PlayCount

- **artist_data.txt** : It provides the names of each artist by their IDs. It has 2 columns separated by tab characters (`\t`).

ArtistID	Name

- **artist_alias.txt**: Note that when plays are scrobbled, the client application submits the name of the artist being played. This name could be misspelled or nonstandard. For example, "The Smiths", "Smiths, The", and "the smiths" may appear as distinct artist IDs in the data set, even though they are plainly the same. `artist_alias.txt` maps artist IDs that are known misspellings or variants to the canonical ID of that artist. The data in this file has 2 columns separated by tab characters (`\t`).

| MisspelledArtistID | StandardArtistID | |---|---|

1.2. Understanding data: simple descriptive statistic

In order to choose or design a suitable algorithm for achieving our goals, given the data we have, we should first understand data characteristics. To start, we import the necessary packages to work with regular expressions, Data Frames, and other nice features of our programming environment.

In [16]:

```
import os
import sys
import re
import random
from pyspark import SparkContext
from pyspark.sql import SQLContext
from pyspark.sql.types import *
from pyspark.sql import Row
from pyspark.sql.functions import *

%matplotlib inline
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from time import time

sqlContext = SQLContext(sc)
base = "/datasets/lastfm/"
```

Question 1

Question 1.0 (Non-grading)

Using SPARK SQL, load data from `/datasets/lastfm/user_artist_data.txt` and show the first 20 entries (via function `show()`).

For this Notebook, from a programming point of view, we are given the schema for the data we use, which is as follows:

```
userID: long int
artistID: long int
playCount: int
```

Each line of the dataset contains the above three fields, separated by a "white space".

In [17]:

```
userArtistDataSchema = StructType([ \
    StructField("userID", LongType(), True), \
    StructField("artistID", LongType(), True), \
    StructField("playCount", IntegerType(), True)])

userArtistDF = sqlContext.read \
    .format('com.databricks.spark.csv') \
    .options(header='false', delimiter=' ') \
    .load(base + "user_artist_data.txt", schema = userArtistDataSchema) \
    .cache()

# we can cache an Dataframe to avoid computing it from the beginning everyt
ime it is accessed.
userArtistDF.cache()

userArtistDF.show()
```

```
+-----+-----+-----+
| userID|artistID|playCount|
+-----+-----+-----+
|1000002|      1|      55|
|1000002| 1000006|      33|
|1000002| 1000007|       8|
|1000002| 1000009|     144|
|1000002| 1000010|     314|
|1000002| 1000013|       8|
|1000002| 1000014|      42|
|1000002| 1000017|      69|
|1000002| 1000024|     329|
|1000002| 1000025|       1|
|1000002| 1000028|      17|
|1000002| 1000031|      47|
|1000002| 1000033|      15|
|1000002| 1000042|       1|
|1000002| 1000045|       1|
|1000002| 1000054|       2|
|1000002| 1000055|      25|
|1000002| 1000056|       4|
|1000002| 1000059|       2|
|1000002| 1000062|      71|
+-----+-----+-----+
only showing top 20 rows
```

Question 1.1:

How many distinct users in data ?

In [18]:

```
uniqueUsers = userArtistDF.select(['userID']).distinct().count()
print("Total number of distinct users: ", uniqueUsers)
```

Total number of distinct users: 148111

Question 1.2

How many distinct artists in data ?

In [19]:

```
uniqueArtists = userArtistDF.select(['artistID']).distinct().count()
print("Total n. of artists: ", uniqueArtists)
```

Total n. of artists: 1631028

We observe that the number of distinct artists is greater than the number of distinct users, so we assume that a user listened to many artists. But this number of artists is still too high probably because of misspelled artists name.

Question 1.3

One limitation of Spark MLlib's ALS implementation - which we will use later- is that it requires IDs for users and items to be nonnegative 32-bit integers. This means that IDs larger than Integer.MAX_VALUE, or 2147483647, can't be used. So we need to check whether this data set conforms to the strict requirements of our library.

What are the maximum and minimum values of column `userID` ?

HINT: Read section 4.3 of Lecture 2 again.

In [20]:

```
userArtistDF.select([min('userID'), max('userID')]).show()
```

```
+-----+-----+
|min(userID)|max(userID)|
+-----+-----+
|          90|    2443548|
+-----+-----+
```

Question 1.4

What are the maximum and minimum values of column `artistID` ?

In [21]:

```
userArtistDF.select([min('artistID'), max('artistID')]).show()
```

```
+-----+-----+
|min(artistID)|max(artistID)|
+-----+-----+
|             1|    10794401|
+-----+-----+
```

We just discovered that we have a total of 148,111 users in our dataset. Similarly, we have a total of 1,631,028 artists in our dataset. The maximum values of `userID` and `artistID` are still smaller than the biggest number of integer type. No additional transformation will be necessary to use these IDs.

One thing we can see here is that SPARK SQL provides us many very concise and powerful tools to do data analytics (comparing to using RDD and their low-level API). You can see more examples [here](#).

Next, we might want to understand better user activity and artist popularity.

Here is a list of simple descriptive queries that helps us reaching these purposes:

- How many times each user has played a song? This is a good indicator of who are the most active users of our service. Note that a very active user with many play counts does not necessarily mean that the user is also "curious"! Indeed, she could have played the same song several times.
- How many play counts for each artist? This is a good indicator of the artist popularity. Since we do not have time information associated to our data, we can only build a, e.g., top-10 ranking of the most popular artists in the dataset. Later in the notebook, we will learn that our dataset has a very "loose" definition about artists: very often artist IDs point to song titles as well. This means we have to be careful when establishing popular artists. Indeed, artists whose data is "well formed" will have the correct number of play counts associated to them. Instead, artists that appear mixed with song titles may see their play counts "diluted" across their songs.

Question 2

Question 2.1

How many times each user has played a song? Show 5 samples of the result.

In [22]:

```
# Compute user activity
# We are interested in how many playcounts each user has scored.
userActivity = userArtistDF.groupBy(['userID']).sum('PlayCount').collect()
print(userActivity[0:5])
print("\n For example the user",userActivity[0][0],"played",userActivity[0][1],"times a song.")
```

```
[Row(userID=1066825, sum(PlayCount)=4), Row(userID=1068108,
sum(PlayCount)=4506), Row(userID=1070733, sum(PlayCount)=14631), Row(userID
=1070807, sum(PlayCount)=3), Row(userID=1071739, sum(PlayCount)=3279)]
```

For example the user 1066825 played 4 times a song.

Question 2.2

Plot CDF (or ECDF) of number of play counts per User ID.

Explain and comment the figure you just created:

- for example, look at important percentiles (25%, median, 75%, tails such as >90%) and cross check with what you have found above to figure out if the result is plausible.
- discuss about your users, with respect to the application domain we target in the notebook: you will notice that for some users, there is very little interaction with the system, which means that maybe recommending something to them is going to be more difficult than for other users who interact more with the system.
- look at outliers and reason about their impact on your recommender algorithm

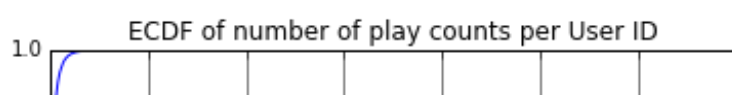
In [51]:

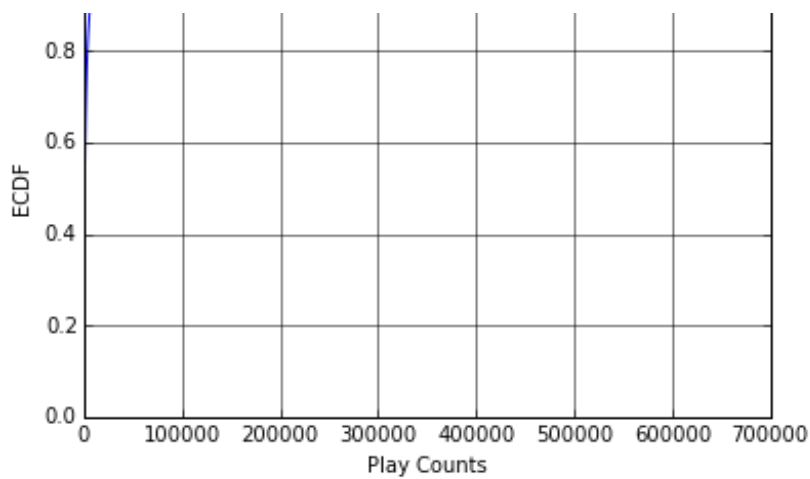
```
pdf = pd.DataFrame(data=userActivity)
Y=np.sort( pdf[1] ) #array of sorted playcount per user
print("The maximum number of playcounts for a single user is " ,np.max(Y))
print('The number of playcounts for the user who is ranked 7000 in the
playcount number is' ,Y[-7000])
yvals=np.arange(len(Y))/float(len(Y))
#Plot only the ECDF of playcounts under 10000 to have a better view of the
data.
Z=[Y[i] for i in range(len(Y)) if Y[i]<10000]
zvals=[yvals[i] for i in range(len(Z))]
#Plot only the ECDF of playcounts under 100000 to see the effect of outlier
s.
W=[Y[i] for i in range(len(Y)) if Y[i]<100000]
wvals=[yvals[i] for i in range(len(W))]
plt.plot( Y, yvals )
plt.xlabel('Play Counts')
plt.ylabel('ECDF')
plt.grid(True,which="both",ls="--")
plt.title('ECDF of number of play counts per User ID')
plt.show()
plt.plot( Z, zvals )
plt.title('ECDF of number of play counts per User ID for a limited number o
f playcounts \n')
plt.show()
plt.plot( W, wvals )
plt.title('ECDF of number of play counts per User ID for a limited number o
f playcounts \n')
plt.show()
p=np.percentile(Y,25)
m=np.percentile(Y,50)
t1=np.percentile(Y,75)
t2=np.percentile(Y,95)
std=np.std(Y)
M=np.mean(Y)

p2=np.percentile(Y,1.24)
print('1.24% of users =',1.24*0.01*len(Y), 'users')
p3=np.percentile(Y,10)
print('10% of users =',10*0.01*len(Y), 'users')
print("1.24% percentile=",p2,"10% percentile=",p3,"25% percentile=",p,"t m
edian=",m,"\n 75% percentile=",t1,"\n tail such as 95%=",t2, "\n the mean="
,M,"\n the standard deviation=",std)
```

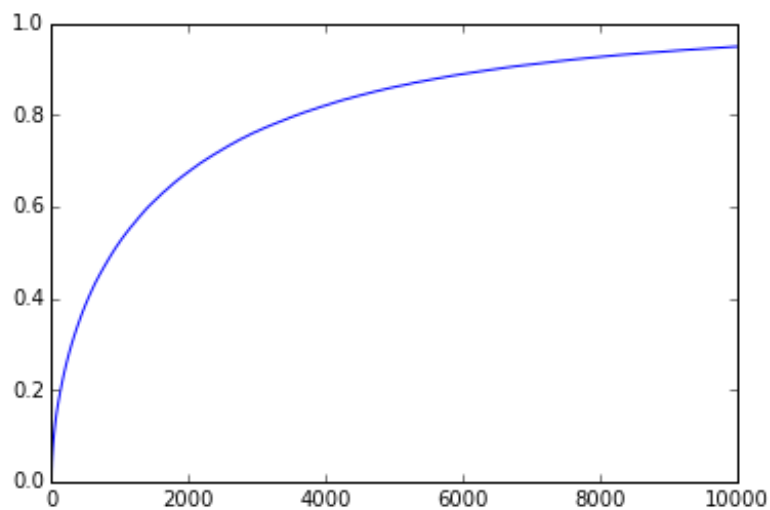
The maximum number of playcounts for a single user is 674412

The number of playcounts for the user who is ranked 7000 in the playcount number is 10457

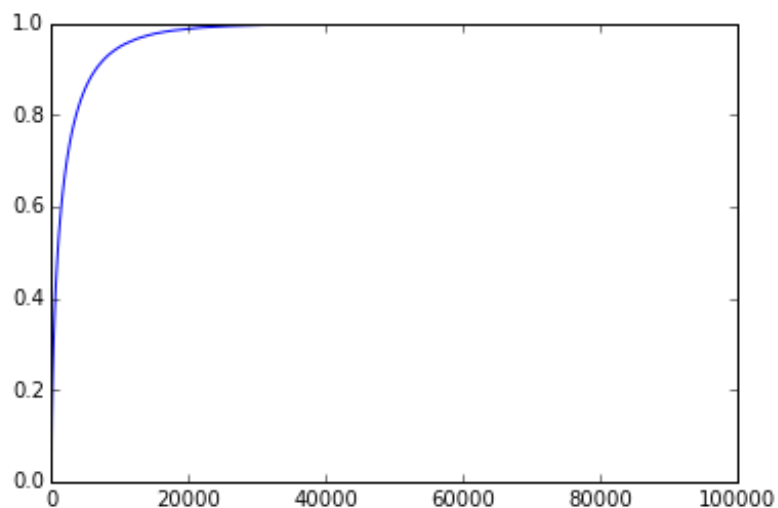




ECDF of number of play counts per User ID for a limited number of playcounts



ECDF of number of play counts per User ID for a limited number of playcounts



1.24% of users = 1836.5764 users

10% of users = 14811.1 users

1.24% percentile= 1.0 10% percentile= 34.0 25% percentile= 204.0 median= 892.0

75% percentile= 2800.0 tail such as 95%= 10120.0

the mean= 2509.1922207

the standard deviation= 5448.29599279

Most users (95%) have listened to a song less than 10000 times. which is plausible because the number of playcounts for the user who is ranked 7000 (which is ~ 5% of the total number of users) in the playcount number is 10457. 1.24% of users (=1825 users) have only listened to one song which

the playcount number is 10000. 10% of users (=14811 users) have only listened to one song which means making recommendations for them quite difficult as they didn't have much interaction with the system. 10% of users (=14811 users) listened to less than 34 songs (not necessarily distinct). Because the lastfm recommender model uses a collaborative filtering technique, it's hard to make recommendations for people who didn't listen to a lot of songs, which means that the more you listen the more you get better recommendations.

The ECDF of number of playcounts per User ID for a number of playcounts less than 10000, comes to confirm what we've been saying, as it climbs to 0.95 for a user with 10000 playcounts, which means that the most active users are very few in comparison with users that are not that active.

We can see from the last plot that the ECDF is almost equal to 1 for a playcount per user greater than 30000, so if we consider outliers the users who listened more than 30000 tracks, we can say that their effects is negligible because they don't have a real impact on the ECDF.

Question 2.3

How many play counts for each artist? Plot CDF or ECDF of the result.

Similarly to the previous question, you need to comment and interpret your result: what is the figure telling you?

In [24]:

```
# Compute artist popularity
# We are interested in how many playcounts per artist
# ATTENTION! Grouping by artistID may be problematic, as stated above.

artistPopularity = userArtistDF.groupby(['artistID']).sum('PlayCount').collect()
artistPopularity[0:5]
```

Out[24]:

```
[Row(artistID=1003514, sum(PlayCount)=949),
 Row(artistID=1004346, sum(PlayCount)=3772),
 Row(artistID=5409, sum(PlayCount)=526693),
 Row(artistID=1002519, sum(PlayCount)=405),
 Row(artistID=1004223, sum(PlayCount)=409)]
```

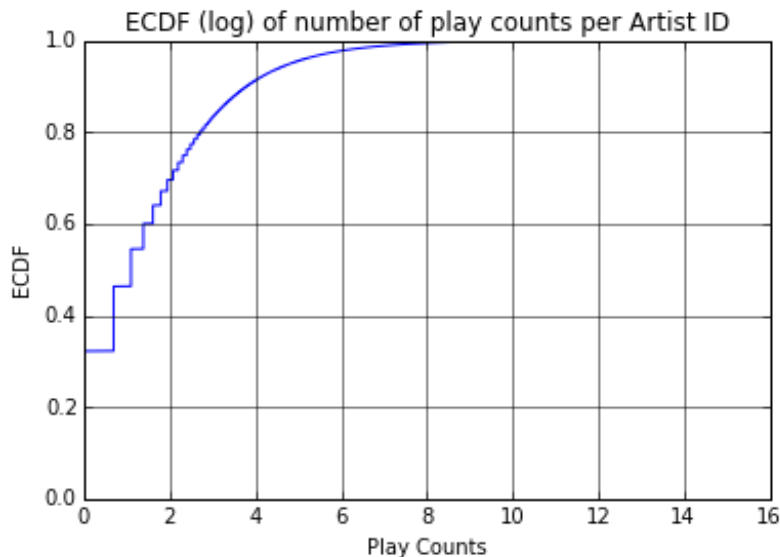
In [25]:

```
#Comment Later
pdf= pd.DataFrame(data=artistPopularity)
Y=np.sort( pdf[1] ) #array of sorted playcount per artist
Y1=Y
Y=np.log(Y)
yvals=np.arange(len(Y))/float(len(Y))
Z=[Y[i] for i in range(len(Y)) if np.exp(Y[i])<126]
zvals=[yvals[i] for i in range(len(Z))]
plt.plot( Y, yvals )
plt.xlabel('Play Counts')
plt.ylabel('ECDF')
plt.grid(True,which="both",ls="--")
plt.title('ECDF (log) of number of play counts per Artist ID')
plt.show()
plt.plot( Z, zvals )
plt.title('ECDF (log) of number of play counts per artist ID for a limited number of playcounts < 126')
```

```

number of playcounts \n )
print('The number of playcounts for the artist who is ranked 81550 in the
playcount number is' ,Y1[-81550])
p=np.percentile(Y1,25)
m=np.percentile(Y1,50)
t1=np.percentile(Y1,75)
t2=np.percentile(Y1,95)
t3=np.percentile(Y1,99)
print("25% percentile=",p,"\\t median=",m,"\\n 75% percentile=",t1,"\\t tail s
uch as 95%=",t2,"\\t tail such as 99%=",t3)

```

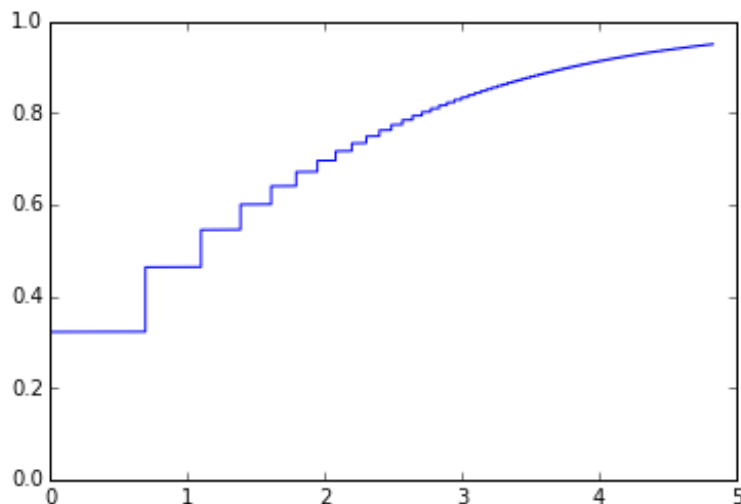


The number of playcounts for the artist who is ranked 81550 in the playcount number is 126

25% percentile= 1.0 median= 3.0

75% percentile= 11.0 tail such as 95%= 126.0 tail such as 99%= 1402.0

ECDF (log) of number of play counts per artist ID for a limited number of playcounts



Most artists (95%) have been listened to less than 126 times. which is plausible because the number of playcounts for the artist who is ranked 81550 (which is ~ 5% of the total number of artists) in the playcount number is 126. 25% of artists have only been listened to one time which means making recommending them nearly impossible as almost nobody listen to them. 95% of artists have only been listened to less than 126 times. Only 1% of the artists have been listened to more than 1400 times. It clearly shows that there is a disproportion in the number of playcounts per artist mainly because a small proportion of artists is very known with a large audience.

The ECDF of number of playcounts per artist ID for a number of playcounts less than 126, comes to

confirm what we've been saying, as it climbs to 0.95 for an artist with 81550 playcounts (~5 in log), which means that the most listened to artists are very few in comparison with artists that are not listened to, that means that most users listen to a small part of artists.

Question 2.4

Plot a bar chart to show top 5 artists In terms of absolute play counts.

Comment the figure you just obtained:

- are these reasonable results?
- is looking at top-5 artists enough to learn more about your data?
- do you see anything strange in the data?

In [26]:

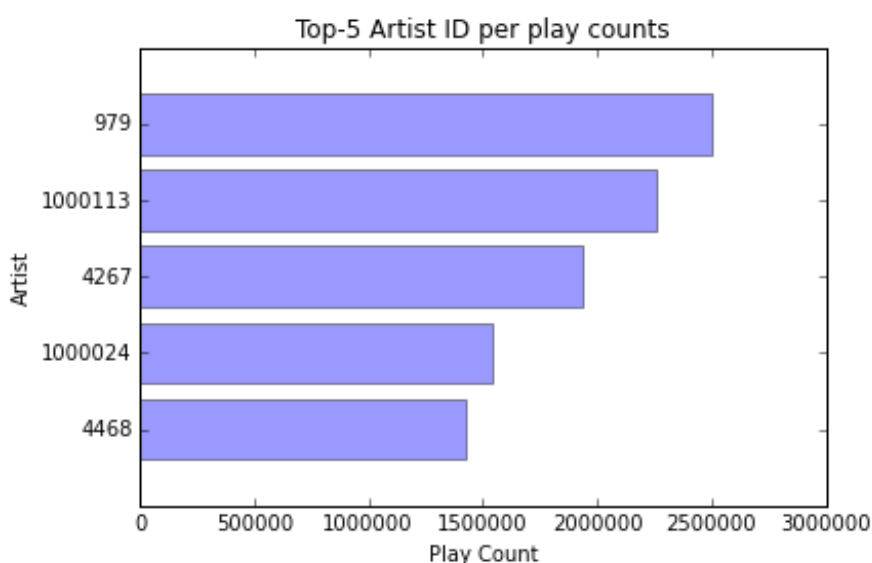
```
sortedArtist = sorted(artistPopularity, key = lambda x: -x[1])[: 5]
print(sortedArtist)
```

```
artistID = [w[0] for w in sortedArtist]
```

```
y_pos = range(len(sortedArtist))
frequency = [w[1] for w in sortedArtist]
```

```
plt.barh(y_pos, frequency[::-1], align='center', alpha=0.4)
plt.yticks(y_pos, artistID[::-1])
plt.xlabel('Play Count')
plt.ylabel('Artist')
plt.title('Top-5 Artist ID per play counts')
plt.show()
```

```
[Row(artistID=979, sum(PlayCount)=2502130), Row(artistID=1000113, sum(PlayC
ount)=2259185), Row(artistID=4267, sum(PlayCount)=1930592),
Row(artistID=1000024, sum(PlayCount)=1542806), Row(artistID=4468, sum(PlayC
ount)=1425942)]
```



the results displayed by these chart are reasonable but not enough to learn more about the data, because we said earlier that most users listen mainly to the first 5% of artists, and these charts don't give us any indication about the other 95% of artists. Moreover, it gives us no information about how many songs each of these artists have. What is also strange about the data is that compared to other

platforms like youtube, the playcount is relatively small, which can make us think of the problem of misspelled artist IDs.

All seems clear right now, but ... wait a second! What about the problems indicated above about artist "disambiguation"? Are these artist ID we are using referring to unique artists? How can we make sure that such "opaque" identifiers point to different bands? Let's try to use some additional dataset to answer this question: `artist_data.txt` dataset. This time, the schema of the dataset consists in:

```
artist ID: long int
name: string
```

We will try to find whether a single singer has two different IDs.

Question 3

Question 3.1

Loading the data from `/datasets/lastfm/artist_data.txt` by using SparkSQL API and show 5 samples.

HINT: If you encounter some error when parsing lines in data because of invalid entries, parameter `mode='DROPMALFORMED'` will help you to eliminate these entries. The suggested syntax is:

```
<df>.options(header='false', delimiter='\t', mode='DROPMALFORMED').
```

In [27]:

```
customSchemaArtist = StructType([ \
    StructField("artistID", LongType(), True), \
    StructField("name", StringType(), True)])

artistDF = sqlContext.read \
    .format('com.databricks.spark.csv') \
    .options(header='false', delimiter='\t', mode='DROPMALFORMED') \
    .load("/datasets/lastfm/artist_data.txt", schema=customSchemaArtist) \
    .cache()

artistDF.show(5)
```

```
+-----+-----+
|artistID|          name|
+-----+-----+
| 1134999|    06Crazy Life|
|  6821360|    Pang Nakarin|
|10113088|Terfel, Bartoli- ...|
|10151459| The Flaming Sidebur|
|  6826647|    Bodenstandig 3000|
+-----+-----+
only showing top 5 rows
```

Question 3.2

Find 20 artists whose name contains "Aerosmith" inside. Take a look on artists that have ID equal to 1000010 and 2082323. In your opinion, are they the same ?

1000010 and 2082323 in your opinion, are they the same?

HINT: Function `locate(sub_string, string)` can be useful in this case.

In [28]:

```
# get artists whose name contains "Aerosmith"
artistDF[locate("Aerosmith", artistDF['name']) > 0].show(20)

# show two examples
artistDF[artistDF.artistID==1000010].show()
artistDF[artistDF.artistID==2082323].show()
```

```
+-----+-----+
|artistID|          name|
+-----+-----+
|10586006|Dusty Springfield...|
| 6946007|    Aerosmith/RunDMC|
|10475683|Aerosmith: Just P...|
| 1083031|    Aerosmith/ G n R|
| 6872848|Britney, Nsync, N...|
|10586963|Green Day - Oasis...|
|10028830|The Aerosmith Ant...|
|10300357| Run-DMC + Aerosmith|
| 2027746|Aerosmith by Musi...|
| 1140418|[rap]Run DMC and ...|
|10237208| Aerosmith + Run DMC|
|10588537|Aerosmith, Kid Ro...|
| 9934757|Aerosmith - Big Ones|
|10437510|Green Day ft. Oas...|
| 6936680| RUN DNC & Aerosmith|
|10479781|    Aerosmith Hits|
|10114147|Charlies Angels -...|
| 1262439|Kid Rock, Run DMC...|
| 7032554|Aerosmith & Run-D...|
|10033592|    Aerosmith?|
+-----+-----+
only showing top 20 rows
```

```
+-----+-----+
|artistID|    name|
+-----+-----+
| 1000010|Aerosmith|
+-----+-----+
```

```
+-----+-----+
|artistID|    name|
+-----+-----+
| 2082323|01 Aerosmith|
+-----+-----+
```

In my opinion, these two artists are the same and we are facing here the problem of non standard/mispelled artists names.

To answer this question correctly, we need to use an additional dataset `artist_alias.txt` which contains the ids of misspelled artists and standard artists. The schema of the dataset consists in:

```
mispelledID ID: long int
```

standard ID: long int

Question 3.3

Using SparkSQL API, load the dataset from `/datasets/lastfm/artist_alias.txt` then show 5 samples.

In [29]:

```
customSchemaArtistAlias = StructType([ \
    StructField("misspelledID", LongType(), True), \
    StructField("standardID", LongType(), True)])

artistAliasDF = sqlContext.read \
    .format('com.databricks.spark.csv') \
    .options(header='false', delimiter='\t', mode='DROPMALFORMED') \
    .load("/datasets/lastfm/artist_alias.txt", schema=customSchemaArtistAlias) \
    .cache()

artistAliasDF.show(5)
```

```
+-----+-----+
|misspelledID|standardID|
+-----+-----+
|    1092764|    1000311|
|    1095122|    1000557|
|    6708070|    1007267|
|   10088054|    1042317|
|    1195917|    1042317|
+-----+-----+
only showing top 5 rows
```

Question 3.4

Verify the answer of question 3.2 ("Are artists that have ID equal to 1000010 and 2082323 the same ?") by finding the standard ids corresponding to the misspelled ids 1000010 and 2082323 respectively.

In [30]:

```
artistAliasDF[artistAliasDF.mispelledID==1000010].show()
artistAliasDF[artistAliasDF.mispelledID==2082323].show()
```

```
+-----+-----+
|misspelledID|standardID|
+-----+-----+

+-----+-----+
|misspelledID|standardID|
+-----+-----+
|    2082323|    1000010|
+-----+-----+
```

1000010 is a standard id, so it haven't been considered as misspelled id in the dataset. 2082323 is a misspelled id and its standard id is 1000010.

Question 4

The misspelled or nonstandard information about artist make our results in the previous queries a bit "sloppy". To overcome this problem, we can replace all misspelled artist ids by the corresponding standard ids and to re-compute the basic descriptive statistics on the "amended" data. First, we construct a "dictionary" that map a non-standard ids to a standard ids. Then this "dictionary" will be used to replace the misspelled artists.

Question 4.1

From data in the dataframe loaded from `/datasets/lastfm/artist_alias.txt`, construct a dictionary that maps each non-standard id to its standard id.

HINT: Instead of using function `collect`, we can use `collectAsMap` to convert the collected data to a dictionary inline.

In [31]:

```
artistAlias = artistAliasDF.rdd.map(lambda row: ( row[0] ,row[1]) ).collectAsMap()

#print(artistAlias)
```

Question 4.2

Using the constructed dictionary in question 4.1, replace the non-standard artist ids in the dataframe that was loaded from `/datasets/lastfm/user_artist_data.txt` by the corresponding standard ids then show 5 samples.

NOTE 1: If an id doesn't exist in the dictionary as a misspelled id, it is really a standard id.

Using function `map` on Spark Dataframe will give us an RDD. We can convert this RDD back to Dataframe by using `sqlContext.createDataFrame(rdd_name, sql_schema)`

NOTE 2: be careful! you need to be able to verify that you indeed solved the problem of having bad artist IDs. In principle, for the new data to be correct, we should to have duplicate pairs (user, artist), potentially with different play counts, right? In answering the question, please **show** that you indeed fixed the problem.

In [32]:

```
from time import time

def replaceMisspelledIDs(fields):
    finalID = artistAlias.get(fields[1], fields[1])
    return (fields[0], finalID, fields[2])

t0 = time()
```

```

newUserArtistDF = sqlContext.createDataFrame(
    userArtistDF.rdd.map(lambda row : replaceMisspelledIDs(row)),
    userArtistDataSchema
)
newUserArtistDF.show(5)

t1 = time()

print('The script takes %f seconds' %(t1-t0))
newUserArtistDF.groupby(['UserID', 'artistID']).count().filter('count>1').show(5)

```

```

+-----+-----+-----+
| userID|artistID|playCount|
+-----+-----+-----+
|1000002|      1|        55|
|1000002| 1000006|        33|
|1000002| 1000007|         8|
|1000002| 1000009|       144|
|1000002| 1000010|       314|
+-----+-----+-----+
only showing top 5 rows

```

The script takes 1.448458 seconds

```

+-----+-----+-----+
| UserID|artistID|count|
+-----+-----+-----+
|1062949| 1000737|     2|
|1062975| 1001819|     2|
|1063004| 2063085|     2|
|1063387|    2536|     2|
|1064168| 6920944|     2|
+-----+-----+-----+
only showing top 5 rows

```

Since there are couples of ('userID','artistID') in the newdatabase that have a count>1, it means that they initially were two different couples: ('userID','misspelledartistID') and ('userID','standardartistID'), which means that the problem is fixed.

Question 4.3

Spark actions are executed through a set of stages, separated by distributed "shuffle" operations. Spark can be instructed to **automatically and efficiently** broadcast common data needed by tasks within **each stage**. The data broadcasted this way is cached in **serialized form** and deserialized before running each task.

We can thus improve our answer to question 4.2: we can reduce the communication cost by shipping the "dictionary" in a more efficient way by using `broadcast variable`. Broadcast variables allow the programmer to keep a read-only variable cached on **each machine** rather than shipping a copy of it with tasks. They are cached in deserialized form. They can be used, for example, to give every node a copy of a large input dataset in an efficient manner.

The broadcast of variable `v` can be created by `bV = sc.broadcast(v)`. Then value of this broadcast variable can be access via `bV.value`

To question is then: using a broadcast variable, modify the script in question 4.2 to get better

No question is there. Using a broadcast variable, modify the script in question 4.2 to get better performance in terms of running time.

In [33]:

```
from time import time

bArtistAlias = sc.broadcast(artistAlias)

def replaceMisspelledIDs(fields):
    finalID = bArtistAlias.value.get(fields[1], fields[1])
    return (fields[0], finalID, fields[2])

t0 = time()

newUserArtistDF = sqlContext.createDataFrame(
    userArtistDF.rdd.map(replaceMisspelledIDs),
    userArtistDataSchema
)
newUserArtistDF.show(5)
t1 = time()

print('The script takes %f seconds' %(t1-t0))
newUserArtistDF = newUserArtistDF.cache()
```

```
+-----+-----+-----+
| userID|artistID|playCount|
+-----+-----+-----+
|1000002|      1|      55|
|1000002| 1000006|      33|
|1000002| 1000007|       8|
|1000002| 1000009|     144|
|1000002| 1000010|     314|
+-----+-----+-----+
only showing top 5 rows
```

The script takes 0.285241 seconds

Using a broadcast variable, the script takes 0.285241 seconds while it was taking 1.448458 seconds without broadcast variable.

Although having some advantages, explicitly creating broadcast variables is only useful when tasks across multiple stages need the same data or when caching the data in deserialized form is important.

Question 5

Well, our data frame contains clean and "standard" data. We can use it to redo previous statistic queries.

Question 5.1

How many unique artists? Compare with the result when using old data.

In [34]:

```
uniqueArtists = newUserArtistDF.select(['artistID']).distinct().count()
```

```
print("New number of artists: ", uniqueArtists, "\n Old number of artists: 1631028")
```

New number of artists: 1568126
Old number of artists: 1631028

Question 5.2

Who are the top-10 artists?

- In terms of absolute play counts
- In terms of "audience size", that is, how many users listened to one of their track at least once

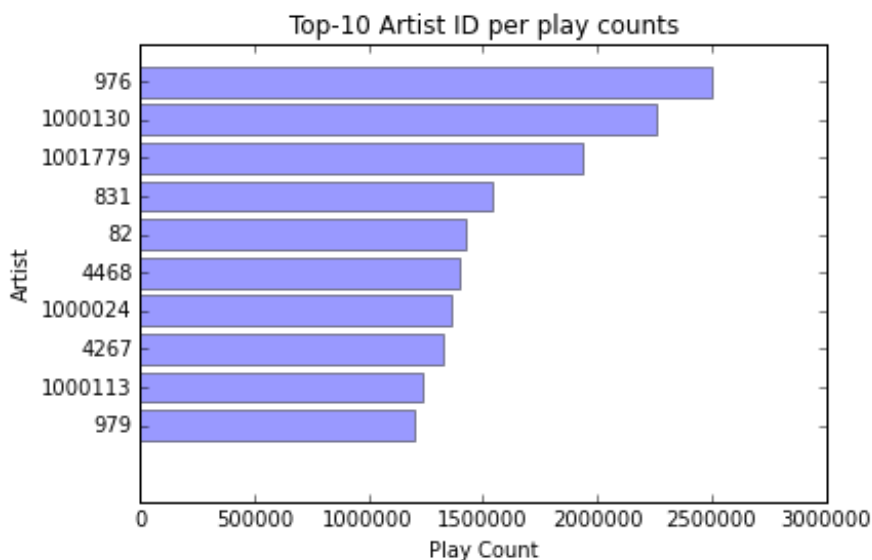
Plot the results, and explain the figures you obtain.

In [35]:

```
# calculate top-10 artists in term of play counts
top10ArtistsPC = newUserArtistDF.groupby(['artistID']).sum('PlayCount').order_by('sum(playCount)', ascending=0).take(10)

y_pos = range(len(top10ArtistsPC))
pdf = pd.DataFrame(data=top10ArtistsPC)

plt.barh(y_pos, pdf[1][::-1], align='center', alpha=0.4)
plt.yticks(y_pos, pdf[0][::-1])
plt.xlabel('Play Count')
plt.ylabel('Artist')
plt.title('Top-10 Artist ID per play counts')
plt.show()
```



If we compare to the results in the questions 2.3 we see that cleaning the data had a major impact in the ranking as some new artists appeared in the ranking and others disappeared. For example the artist 1000130 wasn't in the previous chart but now he's second over all artists.

In [36]:

```
# calculate top-10 artists in term of audience size
top10Artists = newUserArtistDF.select(['userID', 'artistID'])
top10ArtistsAS = top10Artists.drop_duplicates().groupby(['artistID']).count
```

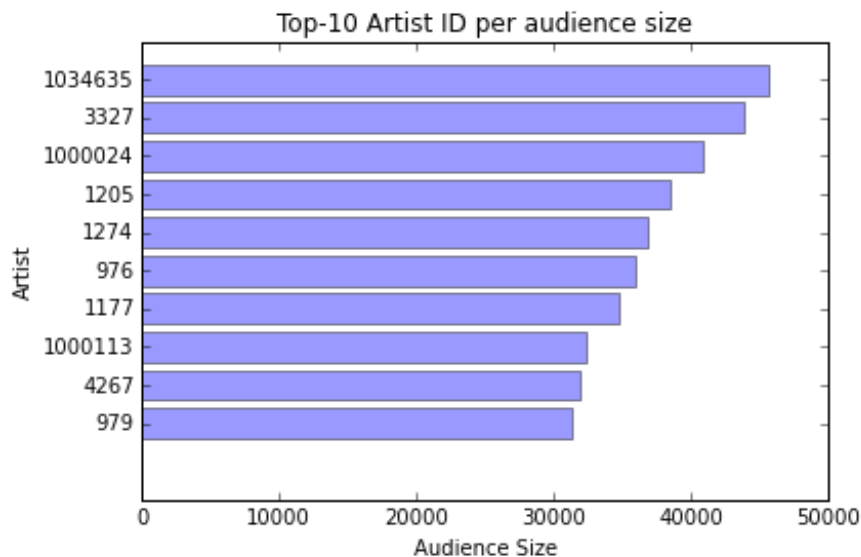
```

top10ArtistsAS = top10ArtistsAS.groupby(['Artist']).count()
().orderBy('count', ascending=0).take(10)

y_pos = range(len(top10ArtistsAS))
pdf = pd.DataFrame(data=top10ArtistsAS)

plt.barh(y_pos, pdf[1][::-1], align='center', alpha=0.4)
plt.yticks(y_pos, pdf[0][::-1])
plt.xlabel('Audience Size')
plt.ylabel('Artist')
plt.title('Top-10 Artist ID per audience size')
plt.show()

```



The ranking is not the same with the previous one as many users can listen to an artist more than one time (which explains why the artist 976 is in first position in number of playcounts but not in audience). For artist 1034635, lots of users listen to him occasionally compared with 976, and that explains why he is not in the first chart.

Question 5.3

Who are the top-10 users?

- In terms of absolute play counts
- In terms of "curiosity", that is, how many different artists they listened to

Plot the results

In [37]:

```

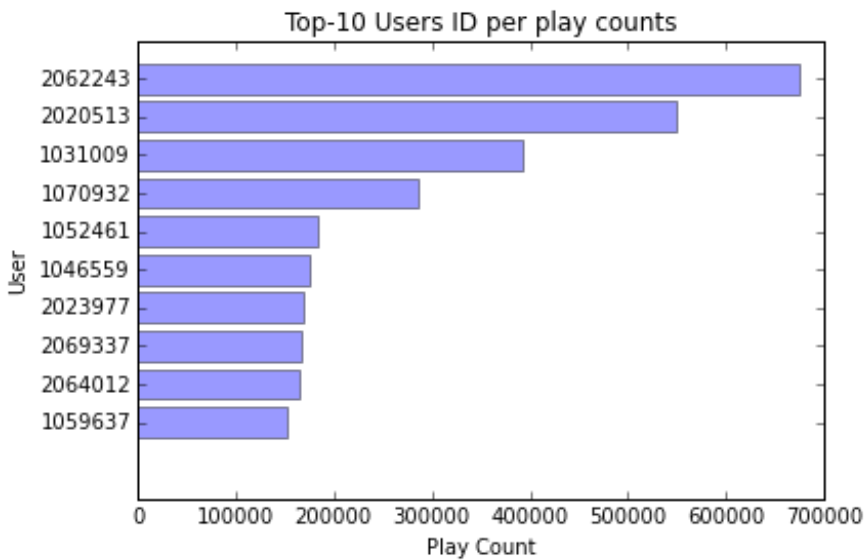
# calculate top 10 users interm of play counts
top10UsersByPlayCount = newUserArtistDF.groupby(['userID']).sum('PlayCount')
().orderBy('sum(PlayCount)', ascending=0).take(10)

y_pos = range(len(top10UsersByPlayCount))
pdf = pd.DataFrame(data=top10UsersByPlayCount)

plt.barh(y_pos, pdf[1][::-1], align='center', alpha=0.4)
plt.yticks(y_pos, pdf[0][::-1])
plt.xlabel('Play Count')
plt.ylabel('User')
plt.title('Top-10 Users ID per play counts')

```

```
plt.show()
```

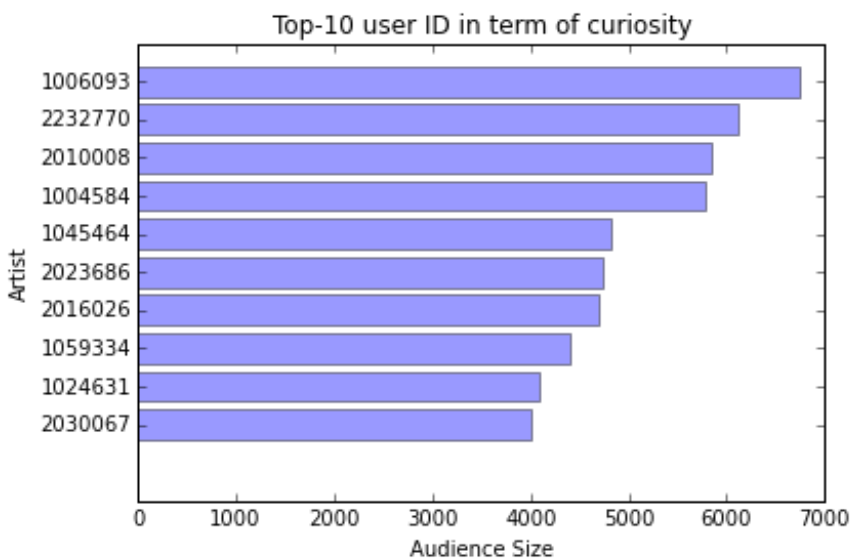


In [38]:

```
# calculate top-10 users in term of curiosity
top10Artists = newUserArtistDF.select(['userID', 'artistID'])
top10ArtistsTC = top10Artists.drop_duplicates().groupBy(['userID']).count().orderBy('count', ascending=0).take(10)

y_pos = range(len(top10ArtistsTC))
pdf = pd.DataFrame(data=top10ArtistsTC)

plt.barh(y_pos, pdf[1][::-1], align='center', alpha=0.4)
plt.yticks(y_pos, pdf[0][::-1])
plt.xlabel('Audience Size')
plt.ylabel('Artist')
plt.title('Top-10 user ID in term of curiosity')
plt.show()
```



Analyzing the difference between the two previous charts gives you an idea about the habit of some users: do they always listen to the same artists or they are more 'curious'.

Now we have some valuable information about the data. It's the time to study how to build a statistical models.

2. Build a statistical models to make recommendations

2.1 Introduction to recommender systems

In a recommendation-system application there are two classes of entities, which we shall refer to as `users` and `items`. Users have preferences for certain items, and these preferences must be inferred from the data. The data itself is represented as a `preference matrix` A , giving for each user-item pair, a value that represents what is known about the degree of preference of that user for that item. The table below is an example for a `preference matrix` of 5 users and k items. The `preference matrix` is also known as `utility matrix`.

	IT1	IT2	IT3	...	ITk
U1	1		5	...	3
U2		2		...	2
U3	5		3	...	
U4	3	3		...	4
U5		1		...	

The value of row i , column j expresses how much does user i like item j . The values are often the rating scores of users for items. An unknown value implies that we have no explicit information about the user's preference for the item. The goal of a recommendation system is to predict "the blanks" in the `preference matrix`. For example, assume that the rating score is from 1 (dislike) to 5 (love), would user $U5$ like $IT3$? We have two approaches:

- Designing our recommendation system to take into account properties of items such as brand, category, price... or even the similarity of their names. We can denote the similarity of items $IT2$ and $IT3$, and then conclude that because user $U5$ did not like $IT2$, they were unlikely to enjoy $IT3$ either.
- We might observe that the people who rated both $IT2$ and $IT3$ tended to give them similar ratings. Thus, we could conclude that user $U5$ would also give $IT3$ a low rating, similar to $U5$'s rating of $IT2$

It is not necessary to predict every blank entry in a `utility matrix`. Rather, it is only necessary to discover some entries in each row that are likely to be high. In most applications, the recommendation system does not offer users a ranking of all items, but rather suggests a few that the user should value highly. It may not even be necessary to find all items with the highest expected ratings, but only to find a large subset of those with the highest ratings.

2.2 Families of recommender systems

In general, recommender systems can be categorized into two groups:

- **Content-Based** systems focus on properties of items. Similarity of items is determined by measuring the similarity in their properties.
- **Collaborative-Filtering** systems focus on the relationship between users and items.

Similarity of items is determined by the similarity of the ratings of those items by the users who have rated both items.

In the usecase of this notebook, artists take the role of `items`, and `users` keep the same role as `users`. Since we have no information about `artists`, except their names, we cannot build a content-based recommender system.

Therefore, in the rest of this notebook, we only focus on Collaborative-Filtering algorithms.

2.3 Collaborative-Filtering

In this section, we study a member of a broad class of algorithms called `latent-factor` models. They try to explain observed interactions between large numbers of users and products through a relatively small number of unobserved, underlying reasons. It is analogous to explaining why millions of people buy a particular few of thousands of possible albums by describing users and albums in terms of tastes for perhaps tens of genres, tastes which are **not directly observable or given** as data.

First, we formulate the learning problem as a matrix completion problem. Then, we will use a type of `matrix factorization` model to "fill in" the blanks. We are given implicit ratings that users have given certain items (that is, the number of times they played a particular artist) and our goal is to predict their ratings for the rest of the items. Formally, if there are n users and m items, we are given an $n \times m$ matrix R in which the generic entry (u, i) represents the rating for item i by user u . **Matrix R has many missing entries indicating unobserved ratings, and our task is to estimate these unobserved ratings.**

A popular approach to the matrix completion problem is **matrix factorization**, where we want to "summarize" users and items with their **latent factors**.

2.3.1 Basic idea and an example of Matrix Factorization

For example, given a preference matrix 5×5 as below, we want to approximate this matrix into the product of two smaller matrixes X and Y .

$$M = \begin{bmatrix} 5 & 2 & 4 & 4 & 3 \\ 3 & 1 & 2 & 4 & 1 \\ 2 & 3 & 1 & 4 & 2 \\ 5 & 4 & 3 & 5 \\ 4 & 4 & 5 & 4 \end{bmatrix} \approx M^{\prime} = \begin{bmatrix} x_{11} & x_{12} & x_{21} & x_{22} \\ x_{31} & x_{32} & x_{41} & x_{42} \\ x_{51} & x_{52} \end{bmatrix} \times \begin{bmatrix} y_{11} & y_{12} & y_{13} & y_{14} & y_{15} \\ y_{21} & y_{22} & y_{23} & y_{24} & y_{25} \end{bmatrix}$$

M^{\prime} is an approximation that is as close to A as possible. To calculate how far from M M^{\prime} is, we often calculate the sum of squared distances of non-empty elements in M and the corresponding elements in M^{\prime} . In this way, for M^{\prime} , besides the approximated elements in M , we also have the non-observed elements. Therefore, to see how much does user i like item j , we simply pick up the value of $M^{\prime}_{i,j}$.

The challenge is how to calculate X and Y . The bad news is that this can't be solved directly for both the best X and best Y at the same time. Fortunately, if Y is known, we can calculate the best of X , and vice versa. It means from the initial values of X and Y in the beginning, we calculate best X according to Y , and then calculate the best Y according to the new X . This process is repeated until the distance from XY to M is converged. It's simple, right ?

Let's take an example. To compute the approximation for the above 5×5 matrix M , first, we init the value of X and Y as below.

$$M^{\prime} = X \times Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}$$

With the initial iteration, we calculate the the Root-Mean-Square Error from XY to M .

Consider the first rows of M and XY . We subtract the first row from XY from the entries in the first row of M , to get $3, 0, 2, 2, 1$. We square and sum these to get 18 .

In the second row, we do the same to get $1, -1, 0, 2, -1$, square and sum to get 7 .

In the third row, the second column is blank, so that entry is ignored when computing the RMSE. The differences are $0, 1, -1, 2$ and the sum of squares is 6 .

For the fourth row, the differences are $0, 3, 2, 1, 3$ and the sum of squares is 23 .

The fifth row has a blank entry in the last column, so the differences are $2, 2, 3, 2$ and the sum of squares is 21 .

When we sum the sums from each of the five rows, we get $18+7+6+23+21 = 75$. So, $RMSE = \sqrt{75/23} = 1.806$ where 23 is the number of non-empty values in M .

Next, with the given value of Y , we calculate X by finding the best value for X_{11} .

$$M^{\prime} = X \times Y = \begin{bmatrix} x & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ x+1 & x+1 & x+1 & x+1 & x+1 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}$$

Now, to minimize the $RMSE$ we minimize the difference of the first rows $(5-(x+1))^2 + (2-(x+1))^2 + (4-(x+1))^2 + (4-(x+1))^2 + (3-(x+1))^2$. By taking the derivative and set that equal to 0, we pick $x=2.6$

Given the new value of X , we can calculate the best value for Y .

$$M^{\prime} = X \times Y = \begin{bmatrix} 2.6 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ y & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 3.6 & 3.6 & 3.6 & 3.6 & 3.6 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{bmatrix}$$

By doing the same process as before, we can pick value for $y=1.617$. After that, we can check if the $RMSE$ is not converged, we continue to update X by Y and vice versa. In this example, for simple, we only update one element of each matrix in each iteration. In practice, we can update a full row or full matrix at once.

2.3.2 Matrix Factorization: Objective and ALS Algorithm on a Single Machine

More formally, in general, we select k latent features, and describe each user u with a k -dimensional vector x_u , and each item i with a k -dimensional vector y_i .

Then, to predict user u 's rating for item i , we do as follows: $r_{ui} \approx x_u^T y_i$.

This can be put, more elegantly, in a matrix form. Let $x_1, \dots, x_n \in \mathbb{R}^k$ be the factors for the users, and $y_1, \dots, y_m \in \mathbb{R}^k$ the factors for the items. The $k \times n$ user matrix X and the $k \times m$ item matrix Y are then defined by:

$$X = \begin{bmatrix} x_1 & \dots & x_n \\ y_1 & \dots & y_i \end{bmatrix} \quad Y = \begin{bmatrix} r_{11} & \dots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{i1} & \dots & r_{in} \end{bmatrix}$$

Our goal is to estimate the complete ratings matrix $R \approx X^T Y$. We can formulate this problem as an optimization problem in which we aim to minimize an objective function and find optimal X and Y . In particular, we aim to minimize the least squares error of the observed ratings (and regularize):

$$\min_{X,Y} \sum_{\{u,i\}} (\text{observed}(r_{ui}) - x_u^T y_i)^2 + \lambda \left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2 \right)$$

Notice that this objective is non-convex (because of the $x_u^T y_i$ term); in fact it's NP-hard to optimize. Gradient descent can be used as an approximate approach here, however it turns out to be slow and costs lots of iterations. Note however, that if we fix the set of variables X and treat them as constants, then the objective is a convex function of Y and vice versa. Our approach will therefore be to fix Y and optimize X , then fix X and optimize Y , and repeat until convergence. This approach is known as **ALS (Alternating Least Squares)**. For our objective function, the alternating least squares algorithm can be expressed with this simple pseudo-code:

Initialize X, Y

while(convergence is not true) **do**

for $u = 1 \dots n$ **do**

$$x_u = \left(\sum_{i \in r_u} y_i y_i^T + \lambda I_k \right)^{-1} \sum_{i \in r_u} r_{ui} y_i$$

end for

for $i = 1 \dots n$ **do**

$$y_i = \left(\sum_{u \in r_i} x_u x_u^T + \lambda I_k \right)^{-1} \sum_{u \in r_i} r_{ui} x_u$$

end for

end while

For a single machine, we can analyze the computational cost of this algorithm. Updating each x_u will cost $O(n_u k^2 + k^3)$, where n_u is the number of items rated by user u , and similarly updating each y_i will cost $O(n_i k^2 + k^3)$, where n_i is the number of users that have rated item i .

Once we've computed the matrices X and Y , there are several ways compute a prediction. The first is to do what was discussed before, which is to simply predict $r_{ui} \approx x_u^T y_i$ for each user u and item i . This approach will cost $O(nmk)$ if we'd like to estimate every user-item pair.

However, this approach is prohibitively expensive for most real-world datasets. A second (and more holistic) approach is to use the x_u and y_i as features in another learning algorithm, incorporating these features with others that are relevant to the prediction task.

2.3.3 Parallel Alternating Least Squares

There are several ways to distribute the computation of the ALS algorithm depending on how data is partitioned.

Method 1: using joins

First we consider a fully distributed version, in the sense that all data (both input and output) is stored in a distributed file system. In practice, input data (ratings) and parameters (X and Y) are stored in an a Spark RDD. Specifically, ratings -- that are always **sparse** -- are stored as RDD of triplets:

Ratings: $\text{RDD}((u, i, r_{ui}), \dots)$

Instead, we can use dense representation for factor matrices X and Y , and these are stored as RDDs of vectors. More precisely, we can use the data types introduced in Spark MLlib to store such vectors and matrices:

$X : \text{RDD}(x_1, \dots, x_n)$

$Y : \text{RDD}(y_1, \dots, y_m)$

Now, recall the expression to compute x_u :

$$x_u = \left(\sum_{i \in r_u} y_i y_i^T + \lambda I_k \right)^{-1} \sum_{i \in r_u} r_{ui} y_i$$

Let's call the first summation *part A* and the second summation *part B*. To compute such parts, in parallel, we can proceed with the following high-level pseudocode:

- Join the Ratings RDD with the Y matrix RDD using key i (items)
- Map to compute $y_i y_i^T$ and emit using key u (user)
- ReduceByKey u (user) to compute $\sum_{i \in r_u} y_i y_i^T$
- Invert
- Another ReduceByKey u (user) to compute $\sum_{i \in r_u} r_{ui} y_i$

We can use the same template to compute y_i .

This approach works fine, but note it requires computing $y_i y_i^T$ for each user that has rated item i .

Method 2: using broadcast variables (advanced topic)

The next approach takes advantage of the fact that the X and Y factor matrices are often very small and can be stored locally on each machine.

- Partition the Ratings RDD **by user** to create R_1 , and similarly partition the Ratings RDD **by item** to create R_2 . This means there are two copies of the same Ratings RDD, albeit with different partitionings. In R_1 , all ratings by the same user are on the same machine, and in R_2 all ratings for same item are on the same machine.
- Broadcast the matrices X and Y . Note that these matrices are not RDD of vectors: they are now "local" matrices.
- Using R_1 and Y , we can use expression x_u from above to compute the update of x_u locally on each machine
- Using R_2 and X , we can use expression y_i from above to compute the update of y_i locally on each machine

A further optimization to this method is to group the X and Y factor matrices into blocks (e.g.,

A further optimization to this method is to group the UX and YR factors matrices into blocks (user blocks and item blocks) and reduce the communication by only sending to each machine the block of users (or items) that are needed to compute the updates at that machine.

This method is called **Block ALS**. It is achieved by precomputing some information about the ratings matrix to determine the "out-links" of each user (which blocks of the items it will contribute to) and "in-link" information for each item (which of the factor vectors it receives from each user block it will depend on). For example, assume that machine 1 is responsible for users 1,2,...,37: these will be block 1 of users. The items rated by these users are block 1 of items. Only the factors of block 1 of users and block 1 of items will be broadcasted to machine 1.

Further readings

Other methods for matrix factorization include:

- Low Rank Approximation and Regression in Input Sparsity Time, by Kenneth L. Clarkson, David P. Woodruff. <http://arxiv.org/abs/1207.6365>
- Generalized Low Rank Models (GLRM), by Madeleine Udell, Corinne Horn, Reza Zadeh, Stephen Boyd. <http://arxiv.org/abs/1410.0342>
- Matrix Completion and Low-Rank SVD via Fast Alternating Least Squares, by Trevor Hastie, Rahul Mazumder, Jason D. Lee, Reza Zadeh. Statistics Department and ICME, Stanford University, 2014. <http://stanford.edu/~rezab/papers/fastals.pdf>

3. Usecase : Music recommender system

In this usecase, we use the data of users and artists in the previous sections to build a statistical model to recommend artists for users.

3.1 Requirements

According to the properties of data, we need to choose a recommender algorithm that is suitable for this implicit feedback data. It means that the algorithm should learn without access to user or artist attributes such as age, genre,... Therefore, an algorithm of type `collaborative filtering` is the best choice.

Second, in the data, there are some users that have listened to only 1 artist. We need an algorithm that might provide decent recommendations to even these users. After all, at some point, every user starts out with just one play at some point!

Third, we need an algorithm that scales, both in its ability to build large models, and to create recommendations quickly. So, an algorithm which can run on a distributed system (SPARK, Hadoop...) is very suitable.

From these requirements, we can choose using ALS algorithm in SPARK's MLLIB.

Spark MLLib's ALS implementation draws on ideas from [1](#) and [2](#).

3.2 Notes

Currently, MLLIB can only build models from an RDD. That means we have two ways to prepare data:

- Loading to into SPARK SQL DataFrame as before, and then access the corresponding RDD by calling `<dataframe>.rdd`. The invalid data is often successfully dropped by using mode `DROPMALFORMED`. However, this way might not work in all cases. Fortunately, we can use it with this usecase.
- Loading data directly to RDD. However, we have to deal with the invalid data ourselves. In the trade-off, this way is the most reliable, and can work in every case.

In this notebook, we will use the second approach: it requires a bit more effort, but the reward is worth it!

3.3 Cleanup the data

In section 1, we already replaced the ids of misspelled artists by the corresponding standard ids by using SPARK SQL API. However, if the data has the invalid entries such that SPARK SQL API is stuck, the best way to work with it is using an RDD.

Just as a recall, we work with three datasets in `user_artist_data.txt`, `artist_alias.txt`. The entries in these files can be empty or have only one field.

In details our goal now is:

- Read the input `user_artist_data.txt` and transform its representation into an output dataset.
- To produce an output "tuple" containing the original user identifier and play counts, but with the artist identifier replaced by its most common alias, as found in the `artist_alias.txt` dataset.
- Since the `artist_alias.txt` file is small, we can use a technique called **broadcast variables** to make such transformation more efficient.

Question 6

Question 6.1

Load data from `/datasets/lastfm/artist_alias.txt` and filter out the invalid entries to construct a dictionary to map from misspelled artists' ids to standard ids.

NOTE: From now on, we will use the "standard" data to train our model.

HINT: If a line contains less than 2 fields or contains invalid numerical values, we can return a special tuple. After that, we can filter out these special tuples.

In [39]:

```
rawArtistAlias = sc.textFile(base + "artist_alias.txt")

def extractFields(s):
    # Using white space or tab character as separators,
    # split a line into list of strings
    line = re.split("\s|\t", s, 1)
    # if this line has at least 2 characters
    if (len(line) > 1):
        try:
            # try to parse the first and the second components to integer t
            v1 = int(line[0])
            v2 = int(line[1])
            return (v1, v2)
```

```

    return (int(line[0]), int(line[1]))
except ValueError:
    # if parsing has any error, return a special tuple
    return (-1,-1)
else:
    # if this line has less than 2 characters, return a special tuple
    return (-1,-1)

artistAlias = (
    rawArtistAlias
    # extract fields using function xtractFields
    .map(lambda row:xtractFields(row))

    # filter out the special tuples
    .filter(lambda row:row[0]!=-1 )

    # collect result to the driver as a "dictionary"
    .collectAsMap()
)

```

Question 6.2

Using the dictionary in question 6.1, prepare RDD `userArtistDataRDD` by replacing misspelled artists' ids to standard ids. Show 5 samples.

HINT: Using broadcast variable can help us increase the efficiency.

In [40]:

```

bArtistAlias = sc.broadcast(artistAlias)
rawUserArtistData = sc.textFile(base + "user_artist_data.txt")

def disambiguate(line):
    [userID, artistID, count] = line.split(' ')
    finalArtistID = bArtistAlias.value.get(artistID,artistID)
    return (userID,finalArtistID, count)

userArtistDataRDD = rawUserArtistData.map(lambda row: disambiguate(row))
userArtistDataRDD.take(5)

```

Out[40]:

```

[('1000002', '1', '55'),
 ('1000002', '1000006', '33'),
 ('1000002', '1000007', '8'),
 ('1000002', '1000009', '144'),
 ('1000002', '1000010', '314')]

```

3.4 Training our statistical model

To train a model using ALS, we must use a preference matrix as an input. MLLIB uses the class `Rating` to support the construction of a distributed preference matrix.

Question 7

Question 7.1

Given RDD `userArtistDataRDD` in question 6.2, construct a new RDD `trainingData` by transforming each item of it into a `Rating` object.

In [41]:

```
from pyspark.mllib.recommendation import ALS, MatrixFactorizationModel, Rating
```

In [42]:

```
allData = userArtistDataRDD.map(lambda r: Rating(r[0], r[1], r[2])).repartition(12).cache()
```

Question 7.2

A model can be trained by using `ALS.trainImplicit(<training data>, <rank>)`, where:

- `training data` is the input data you decide to feed to the ALS algorithm
- `rank` is the number of latent features

We can also use some additional parameters to adjust the quality of the model. Currently, let's set

- `rank=10`
- `iterations=5`
- `lambda_=0.01`
- `alpha=1.0`

to build model.

In [52]:

```
t0 = time()
model = ALS.trainImplicit(allData, 10)
t1 = time()
print("finish training model in %f secs" % (t1 - t0))
```

finish training model in 38.646748 secs

Question 7.3

The trained model can be saved into HDFS for later use. This can be done via `model.save(sc, <file_name>)`. Let's use this function to store our model as name `lastfm_model.spark`.

NOTE 1: since you may have noticed that building the model takes some time, it might come to your mind that this information could be stored, such that you can "interrupt" your laboratory session here, and restart next time by loading your model.

NOTE 2: funnily enough, it could take more time to save the model than to build it from scratch! So take a look at the execution time to save the model: this method actually stores the model as Parquet files, which are column-oriented and compressed.

NOTE 3: to check you have your file on HDFS, you are invited to open a terminal from the "Home" Jupyter dashboard, and type `hdfs dfs -ls` to check.

In [44]:

```
! hdfs dfs -rm -R -f -skipTrash lastfm_model.spark  
model.save(sc, 'lastfm_model.spark')
```

Deleted lastfm_model.spark

Question 7.4

A saved model can be load from file by using `MatrixFactorizationModel.load(sc, <file_name>)`.

Let's load our model from file.

In [45]:

```
t0 = time()  
model = MatrixFactorizationModel.load(sc, 'lastfm_model.spark')  
t1 = time()  
print("finish loading model in %f secs" % (t1 - t0))
```

finish loading model in 1.620845 secs

Question 7.5

Print the first row of user features in our model.

In [46]:

```
model.userFeatures().take(1)
```

Out[46]:

```
[(120,  
  array('d', [0.02505505457520485, 0.04512013867497444,  
0.03700440004467964, -0.03723803907632828, -0.028402971103787422, -0.004354  
292061179876, -0.005954351741820574, 0.03587596118450165, -  
0.010521704331040382, 0.0006468526553362608]))]
```

Question 8

Show the top-5 artist names recommended for user 2093760.

HINT: The recommendations can be given by function `recommendProducts(userID, num_recommendations)`. These recommendations are only artist ids. You have to map them to artist names by using data in `artist_data.txt`.

In [36]:

```
# Make five recommendations to user 2093760  
recommendations = (model.recommendProducts(2093760,5))  
  
# construct set of recommended artists  
recArtist = set(recommendations)
```

In [37]:

```
# construct data of artists (artist_id, artist_name)

rawArtistData = sc.textFile(base + "artist_data.txt")

def xtractFields(s):
    line = re.split("\s|\t",s,1)
    if (len(line) > 1):
        try:
            return (int(line[0]), str(line[1].strip()))
        except ValueError:
            return (-1, "")
    else:
        return (-1, "")


artistByID = rawArtistData.map(xtractFields).filter(lambda x: x[0] > 0)
```

In [38]:

```
# Filter in those artists, get just artist, and print
def artistNames(line):
    # [artistID, name]
    if (line[0] in (recommendations[i][1] for i in range (len(recommendation
s)))):
        return True
    else:
        return False

recList = artistByID.filter(artistNames).values().collect()

print(recList)
```



```
['50 Cent', 'Snoop Dogg', 'Nas', 'Jay-Z', 'Outkast']
```

IMPORTANT NOTE

At the moment, it is necessary to manually unpersist the RDDs inside the model when you are done with it. The following function can be used to make sure models are promptly uncached.

In [47]:

```
def unpersist(model):
    model.userFeatures().unpersist()
    model.productFeatures().unpersist()

# uncache data and model when they are no longer used
unpersist(model)
```

3.5 Evaluating Recommendation Quality

In this section, we study how to evaluate the quality of our model. It's hard to say how good the recommendations are. One of several methods approach to evaluate a recommender based on its ability to rank good items (artists) high in a list of recommendations. The problem is how to define "good artists". Currently, by training all data, "good artists" is defined as "artists the user has listened to", and the recommender system has already received all of this information as input. It could trivially

return the users previously-listened artists as top recommendations and score perfectly. Indeed, this is not useful, because the recommender's is used to recommend artists that the user has **never** listened to.

To overcome that problem, we can hide some of the artist play data and only use the rest to train model. Then, this held-out data can be interpreted as a collection of "good" recommendations for each user. The recommender is asked to rank all items in the model, and the rank of the held-out artists are examined. Ideally the recommender places all of them at or near the top of the list.

The recommender's score can then be computed by comparing all held-out artists' ranks to the rest. The fraction of pairs where the held-out artist is ranked higher is its score. 1.0 is perfect, 0.0 is the worst possible score, and 0.5 is the expected value achieved from randomly ranking artists.

AUC(Area Under the Curve) can be used as a metric to evaluate model. It is also viewed as the probability that a randomly-chosen "good" artist ranks above a randomly-chosen "bad" artist.

Next, we split the training data into 2 parts: `trainData` and `cvData` with ratio 0.9:0.1 respectively, where `trainData` is the dataset that will be used to train model. Then we write a function to calculate AUC to evaluate the quality of our model.

Question 9

Question 9.1

Split the data into `trainData` and `cvData` with ratio 0.9:0.1 and use the first part to train a statistic model with:

- rank=10
- iterations=5
- lambda_=0.01
- alpha=1.0

In [48]:

```
trainData, cvData = allData.randomSplit([0.9,0.1]) #splitting data into 2 parts : training and validation
trainData.cache()
cvData.cache()
```

Out[48]:

PythonRDD[360] at RDD at PythonRDD.scala:48

In [54]:

```
t0 = time()
#Using training data to train the model

model = ALS.trainImplicit(trainData,10,5,0.01,1)
t1 = time()
print("finish training model in %f secs" % (t1 - t0))
```

finish training model in 102.177056 secs

Area under the ROC curve: a function to compute it

In [45]:

```
# Get all unique artistId, and broadcast them
allItemIDs = np.array(allData.map(lambda x: x[1]).distinct().collect())
bAllItemIDs = sc.broadcast(allItemIDs)
```

In [46]:

```
from random import randint

# Depend on the number of item in userIDAndPosItemIDs,
# create a set of "negative" products for each user. These are randomly
chosen
# from among all of the other items, excluding those that are "positive" for
the user.
# NOTE 1: mapPartitions operates on many (user,positive-items) pairs at once
# NOTE 2: flatMap breaks the collections above down into one big set of
tuples
def xtractNegative(userIDAndPosItemIDs):
    def pickEnoughNegatives(line):
        userID = line[0]
        posItemIDSet = set(line[1])
        #posItemIDSet = line[1]
        negative = []
        allItemIDs = bAllItemIDs.value
        # Keep about as many negative examples per user as positive. Duplicates
        are OK.
        i = 0
        while (i < len(allItemIDs) and len(negative) < len(posItemIDSet)):
            itemID = allItemIDs[randint(0,len(allItemIDs)-1)]
            if itemID not in posItemIDSet:
                negative.append(itemID)
            i += 1

        # Result is a collection of (user,negative-item) tuples
        return map(lambda itemID: (userID, itemID), negative)

    # Init an RNG and the item IDs set once for partition
    # allItemIDs = bAllItemIDs.value
    return map(pickEnoughNegatives, userIDAndPosItemIDs)

def ratioOfCorrectRanks(positiveRatings, negativeRatings):

    # find number elements in arr that has index >= start and has value smaller
    than x
    # arr is a sorted array
    def findNumElementsSmallerThan(arr, x, start=0):
        left = start
        right = len(arr) - 1
        # if x is bigger than the biggest element in arr
        if start > right or x > arr[right]:
            return right + 1
        mid = -1
        while left <= right:
            mid = (left + right) // 2
            if arr[mid] < x:
                left = mid + 1
            elif arr[mid] > x:
                right = mid - 1
```

```

        while mid-1 >= start and arr[mid-1] == x:
            mid -= 1
        return mid
    return mid if arr[mid] > x else mid + 1

    ## AUC may be viewed as the probability that a random positive item
scores
    ## higher than a random negative one. Here the proportion of all posi
ve-negative
    ## pairs that are correctly ranked is computed. The result is equal to
the AUC metric.
    correct = 0 ## L
    total = 0 ## L

    # sorting positiveRatings array needs more cost
    #positiveRatings = np.array(map(lambda x: x.rating, positiveRatings))

    negativeRatings = list(map(lambda x:x.rating, negativeRatings))

    #np.sort(positiveRatings)
    negativeRatings.sort() # = np.sort(negativeRatings)
    total = len(positiveRatings)*len(negativeRatings)

    for positive in positiveRatings:
        # Count the correctly-ranked pairs
        correct += findNumElementsSmallerThan(negativeRatings, positive.rati
ing)

    ## Return AUC: fraction of pairs ranked correctly
    return float(correct) / total

def calculateAUC(positiveData, bAllItemIDs, predictFunction):
    # Take held-out data as the "positive", and map to tuples
    positiveUserProducts = positiveData.map(lambda r: (r[0], r[1]))
    # Make predictions for each of them, including a numeric score, and gat
her by user
    positivePredictions = predictFunction(positiveUserProducts).groupBy(lam
bda r: r.user)

    # Create a set of "negative" products for each user. These are randomly
chosen
    # from among all of the other items, excluding those that are "positive
" for the user.
    negativeUserProducts = positiveUserProducts.groupByKey().mapPartitions(
extractNegative).flatMap(lambda x: x)
    # Make predictions on the rest
    negativePredictions = predictFunction(negativeUserProducts).groupBy(lam
bda r: r.user)

    return (
        positivePredictions.join(negativePredictions)
            .values()
            .map(
                lambda positive_negativeRatings: ratioOfCorrectRanks(pos
itive_negativeRatings[0], positive_negativeRatings[1])
            )
            .mean()
    )

```

Question 9.2

Using part `cvData` and function `calculateAUC` to compute the AUC of the trained model.

In [47]:

```
t0 = time()
auc = calculateAUC(cvData,bAllItemIDs, model.predictAll)
t1 = time()
print("auc=",auc)
print("finish in %f seconds" % (t1 - t0))
```

```
auc= 0.9633206447520896
finish in 25.630683 seconds
```

Question 9.3

Now we have the UAC of our model, it's helpful to benchmark this against a simpler approach. For example, consider recommending the globally most-played artists to every user. This is not personalized, but is simple and may be effective.

Implement this simple popularity-based prediction algorithm, evaluate its AUC score, and compare to the results achieved by the more sophisticated ALS algorithm.

In [48]:

```
bListenCount = sc.broadcast(trainData.map(lambda r: (r[1],
r[2])).reduceByKey(lambda f1,f2:f1+f2 ).collectAsMap())
def predictMostListened(allData):
    return allData.map(lambda r: Rating(r[0], r[1], bListenCount.value.get(
r[1] , 0.0)))
```

In [49]:

```
auc = calculateAUC(cvData,bAllItemIDs, predictMostListened)
print(auc)
```

```
0.9372581714312582
```

3.6 Personalized recommendations with ALS

In the previous section, we build our models with some given parameters without any knowledge about them. Actually, choosing the best parameters' values is very important. It can significantly affect the quality of models. Especially, with the current implementation of ALS in MLLIB, these parameters are not learned by the algorithm, and must be chosen by the caller. The following parameters should get consideration before training models:

- `rank = 10`: the number of latent factors in the model, or equivalently, the number of columns k in the user-feature and product-feature matrices. In non-trivial cases, this is also their rank.
- `iterations = 5`: the number of iterations that the factorization runs. Instead of running the algorithm until RMSE converged which actually takes very long time to finish with large datasets, we only let it run in a given number of iterations. More iterations take more time

but may produce a better factorization.

- `lambda_ = 0.01`: a standard overfitting parameter. Higher values resist overfitting, but values that are too high hurt the factorization's accuracy.
- `alpha = 1.0`: controls the relative weight of observed versus unobserved userproduct interactions in the factorization.

Although all of them have impact on the models' quality, `iterations` is more of a constraint on resources used in the factorization. So, `rank`, `lambda_` and `alpha` can be considered hyperparameters to the model. We will try to find "good" values for them. Indeed, the values of hyperparameter are not necessarily optimal. Choosing good hyperparameter values is a common problem in machine learning. The most basic way to choose values is to simply try combinations of values and evaluate a metric for each of them, and choose the combination that produces the best value of the metric.

Question 10

Question 10.1

For simplicity, assume that we want to explore the following parameter space: `rank` in `{10, 50}`, `lambda_` in `{1.0, 0.0001}` and `alpha` in `{1.0, 40.0}`.

Find the best combination of them in terms of the highest AUC value.

In [50]:

```
evaluations = []

for rank in [10, 50]:
    for lambda_ in [1.0, 0.0001]:
        for alpha in [1, 40]:
            print("Train model with rank=%d lambda_=%f alpha=%f" % (rank,
lambda_, alpha))
            # with each combination of params, we should run multiple times
and get avg
            # for simple, we only run one time.
            model = ALS.trainImplicit(trainData,rank,5,lambda_,alpha)

            auc = calculateAUC(cvData,bAllItemIDs,model.predictAll)

            evaluations.append((rank, lambda_, alpha), auc))

            unpersist(model)

evaluations.sort(key=lambda x: -x[-1])

evalDataFrame = pd.DataFrame(data=evaluations)
print(evalDataFrame)

trainData.unpersist()
cvData.unpersist()
```

```
Train model with rank=10 lambda_=1.000000 alpha=1.000000
Train model with rank=10 lambda_=1.000000 alpha=40.000000
Train model with rank=10 lambda_=0.000100 alpha=1.000000
Train model with rank=10 lambda_=0.000100 alpha=40.000000
Train model with rank=50 lambda_=1.000000 alpha=1.000000
```

```

Train model with rank=50 lambda_=1.000000 alpha=40.000000
Train model with rank=50 lambda_=0.000100 alpha=1.000000
Train model with rank=50 lambda_=0.000100 alpha=40.000000

```

	0	1
0	(50, 1.0, 40)	0.974191
1	(50, 1.0, 1)	0.974010
2	(10, 1.0, 40)	0.971217
3	(10, 1.0, 1)	0.970718
4	(10, 0.0001, 1)	0.962361
5	(10, 0.0001, 40)	0.961637
6	(50, 0.0001, 1)	0.940900
7	(50, 0.0001, 40)	0.940673

Out[50]:

PythonRDD[492] at RDD at PythonRDD.scala:48

Question 10.2

Using "optimal" hyper-parameters in question 10.1, re-train the model and show top-5 artist names recommended for user 2093760.

In [51]:

```

model = ALS.trainImplicit(trainData,50,5,1.0,40)
allData.unpersist()

userID = 2093760
recommendations = model.recommendProducts(2093760,5)

recommendedProductIDs = set(recommendations )

recList = artistByID.filter(artistNames).values().collect()
print(recList)

unpersist(model)

```

```

['The Nightmare Scenario', 'Whatever It Takes', 'Peter Brame', 'Josh Rosenb
lum', 'Punk Goes Acoustic']

```

Summary

In this notebook, we introduce an algorithm to do matrix factorization and the way of using it to make recommendation. Further more, we studied how to build a large-scale recommender system on SPARK using ALS algorithm and evaluate its quality. Finally, a simple approach to choose good parameters is mentioned.

References

- The example in section 2 is taken from [Recommender system](#)