Music Recommender System using Apache Spark and Python

Estimated time: 8hrs

Description

For this project, you are to create a recommender system that will recommend new musical artists to a user based on their listening history. Suggesting different songs or musical artists to a user is important to many music streaming services, such as Pandora and Spotify. In addition, this type of recommender system could also be used as a means of suggesting TV shows or movies to a user (e.g., Netflix).

To create this system you will be using Spark and the collaborative filtering technique. The instructions for completing this project will be laid out entirely in this file. You will have to implement any missing code as well as answer any questions.

Submission Instructions:

- Add all of your updates to this IPython file and do not clear any of the output you get from running your code.
- Upload this file onto moodle.

Datasets

You will be using some publicly available song data from audioscrobbler, which can be found here (http://www-etud.iro.umontreal.ca/~bergstrj/audioscrobbler_data.html). However, we modified the original data files so that the code will run in a reasonable time on a single machine. The reduced data files have been suffixed with `_small.txt` and contains only the information relevant to the top 50 most prolific users (highest artist play counts).

The original data file `user_artist_data.txt` contained 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.

Note that when plays are scribbled, the client application submits the name of the artist being played. This name could be misspelled or nonstandard, and this may only be detected later. For example, "The Smiths", "Smiths, The", and "the smiths" may appear as distinct artist IDs in the data set, even though they clearly refer to the same artist. So, the data set includes `artist_alias.txt`, which maps artist IDs that are known misspellings or variants to the canonical ID of that artist.

The `artist_data.txt` file then provides a map from the canonical artist ID to the name of the artist.

Necessary Package Imports

In [1]:
```
from pyspark.mllib.recommendation import *
import random
from operator import *
```

Loading data

Load the three datasets into RDDs and name them `artistData`, `artistAlias`, and `userArtistData`. View the README, or the files themselves, to see how this data is formated. Some of the files have tab delimeters while some have space delimiters. Make sure that your `userArtistData` RDD contains only the canonical artist IDs.

In [2]:
```
artistData=sc.textFile("/home/shriyansh/artist_data_small.txt")
artistData = artistData.map(lambda x : x.split("\t")).map(lambda x: [int(x[0])

artistAlias=sc.textFile("/home/shriyansh/artist_alias_small.txt").map(lambda x
artistAlias=artistAlias.map(lambda x: [int(x[0]),int(x[1])]).collectAsMap()

userArtistData=sc.textFile("/home/shriyansh/user_artist_data_small.txt").map(l
userArtistData=userArtistData.map(lambda x: [int(x[0]),int(x[1]),int(x[2])])

for data in userArtistData.collect() :
        if data[1] in artistAlias:
                data = (data[0], artistAlias[data[1]], data[2])
                break
```

Data Exploration

In the blank below, write some code that with find the users' total play counts. Find the three users with the highest number of total play counts (sum of all counters) and print the user ID, the total play count, and the mean play count (average number of times a user played an artist). Your output should look as follows:

```
User 1059637 has a total play count of 674412 and a mean play count of
1878.
User 2064012 has a total play count of 548427 and a mean play count of
9455.
User 2069337 has a total play count of 393515 and a mean play count of
1519.
```

In [3]:
```
#playcount={}
#count={}
#for user in userArtistData:
#   playcount[user]=userArtistData.groupBy()

from collections import OrderedDict
uad={}
count={}
mean={}
for data in userArtistData.collect() :
    if not data[0] in uad.keys():
        uad[data[0]] = int(data[2])
        count[data[0]] = 1
    else:
        uad[data[0]]=uad[data[0]]+data[2]
        count[data[0]] = count[data[0]] + 1

#print len(uad)
for i in uad.keys():
    mean[i]=uad[i]/count[i]

sorted_uad=OrderedDict(sorted(uad.items(),key=lambda kv:kv[1], reverse=True))

number=0
for user in sorted_uad.keys():
    print "User ", user , " has a total play count of ", uad[user], "and a mea
    number+=1
    if number==3:
        break
```

```
User  1059637  has a total play count of  674412 and a mean play count of  18
78
User  2064012  has a total play count of  548427 and a mean play count of  94
55
User  2069337  has a total play count of  393515 and a mean play count of  15
19
```

Splitting Data for Testing

Use the randomSplit (http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD.randomSplit) function to divide the data (`userArtistData`) into:

- A training set, `trainData`, that will be used to train the model. This set should constitute 40% of the data.
- A validation set, `validationData`, used to perform parameter tuning. This set should constitute 40% of the data.
- A test set, `testData`, used for a final evaluation of the model. This set should constitute 20% of the data.

Use a random seed value of 13. Since these datasets will be repeatedly used you will probably want to persist them in memory using the cache (http://spark.apache.org/docs/latest/api/python/pyspark.html#pyspark.RDD.cache) function.

In addition, print out the first 3 elements of each set as well as their sizes; if you created these sets correctly, your output should look as follows:

```
[(1059637, 1000049, 1), (1059637, 1000056, 1), (1059637, 1000113, 5)]
[(1059637, 1000010, 238), (1059637, 1000062, 11), (1059637, 1000112, 42
3)]
[(1059637, 1000094, 1), (1059637, 1000130, 19129), (1059637, 1000139, 4
)]
19817
19633
10031
```

In [4]:
```
trainData,validationData,testData=userArtistData.randomSplit([40,40,20],13)

trainData.cache()
validationData.cache()
testData.cache()

print trainData.take(3)
print validationData.take(3)
print testData.take(3)

print trainData.count()
print validationData.count()
print testData.count()
```

```
[[1059637, 1000049, 1], [1059637, 1000056, 1], [1059637, 1000113, 5]]
[[1059637, 1000010, 238], [1059637, 1000062, 11], [1059637, 1000112, 423]]
[[1059637, 1000094, 1], [1059637, 1000130, 19129], [1059637, 1000139, 4]]
19817
19633
10031
```

The Recommender Model

For this project, we will train the model with implicit feedback. You can read more information about this from the collaborative filtering page: http://spark.apache.org/docs/latest/mllib-collaborative-filtering.html. The [function you will be using](http://spark.apache.org/docs/latest/api/python/pyspark.mllib.html#pyspark.mllib.recommendation.ALS.trainImplicit) has a few tunable parameters that will affect how the model is built. Therefore, to get the best model, we will do a small parameter sweep and choose the model that performs the best on the validation set

Therefore, we must first devise a way to evaluate models. Once we have a method for evaluation, we can run a parameter sweep, evaluate each combination of parameters on the validation data, and choose the optimal set of parameters. The parameters then can be used to make predictions on the test data.

Model Evaluation

Although there may be several ways to evaluate a model, we will use a simple method here. Suppose we have a model and some dataset of *_true_* artist plays for a set of users. This model can be used to predict the top X artist recommendations for a user and these recommendations can be compared the artists that the user actually listened to (here, X will be the number of artists in the dataset of *_true_* artist plays). Then, the fraction of overlap between the top X predictions of the model and the X artists that the user actually listened to can be calculated. This process can be repeated for all users and an average value returned.

For example, suppose a model predicted [1,2,4,8] as the top X=4 artists for a user. Suppose, that user actually listened to the artists [1,3,7,8]. Then, for this user, the model would have a score of 2/4=0.5. To get the overall score, this would be performed for all users, with the average returned.

****NOTE: when using the model to predict the top-X artists for a user, do not include the artists listed with that user in the training data.****

Name your function `modelEval` and have it take a model (the output of ALS.trainImplicit) and a dataset as input. For parameter tuning, the dataset parameter should be set to the validation data (`validationData`). After parameter tuning, the model can be evaluated on the test data (`testData`).

In [5]:
```
def modelEval(model,dataset):
    #Step 1:Finding all the artists
    artists=userArtistData.map(lambda x:x[1]).distinct()

    #Step 2: Finding all the users from the considered dataset
    users=dataset.map(lambda x:x[0]).distinct().collect()

    model_score=0.000000
    u_count=len(users)

    for user in users:
    #Step 3: Finding the artists similarity and overall score
        data_artists=dataset.filter(lambda x:x[0]==user).map(lambda x:x[1])
        training_artists=trainData.filter(lambda x:x[0]==user).map(lambda x:x[
        remaining_artists=artists.filter(lambda x: x not in training_artists)

        remaining_artists_users=remaining_artists.map(lambda x:(user,x))

        count=len(data_artists.collect())

        predictions=model.predictAll(remaining_artists_users)
        sorted_predictions=predictions.takeOrdered(count,key=lambda x:-x[2])
        sorted_predictions_RDD=sc.parallelize(sorted_predictions)

        req_predictions=sorted_predictions_RDD.map(lambda x: x[1])

        similarity=data_artists.intersection(req_predictions)
        s_count=len(similarity.collect())

        output=float(s_count)/float(count)
        model_score= model_score + output

    overall_score=float(model_score)/float(u_count)
    #print overall_score

    return overall_score
```

Model Construction

Now we can build the best model possibly using the validation set of data and the `modelEval` function. Although, there are a few parameters we could optimize, for the sake of time, we will just try a few different values for the [rank parameter](http://spark.apache.org/docs/latest/mllib-collaborative-filtering.html#collaborative-filtering) (leave everything else at its default value, **except make `seed`=345**). Loop through the values [2, 10, 20] and figure out which one produces the highest scored based on your model evaluation function.

Note: this procedure may take several minutes to run.

For each rank value, print out the output of the `modelEval` function for that model. Your output should look as follows:

```
The model score for rank 2 is 0.090431
The model score for rank 10 is 0.095294
The model score for rank 20 is 0.090248
```

In [6]:
```
ranks=[2,10,20]
for value in ranks:
    model = ALS.trainImplicit(trainData, rank= value, seed=345)
    output = modelEval(model, validationData)
    print "The model score for rank "+str(value)+" is "+str(output)
```

```
The model score for rank 2 is 0.0928696790111
The model score for rank 10 is 0.0972298276899
The model score for rank 20 is 0.0832988044617
```

Now, using the bestModel, we will check the results over the test data. Your result should be ~`0.0507`.

In [7]:
```
bestModel = ALS.trainImplicit(trainData, rank=10, seed=345)
modelEval(bestModel, testData)
```

Out[7]:
```
0.05893319359387068
```

Trying Some Artist Recommendations

Using the best model above, predict the top 5 artists for user `1059637` using the recommendProducts (http://spark.apache.org/docs/1.5.2/api/python/pyspark.mllib.html#pyspark.mllib.recommendation.MatrixFactorizationModel.recommendProducts) function. Map the results (integer IDs) into the real artist name using `artistAlias`. Print the results. The output should look as follows:

```
Artist 0: Brand New
Artist 1: Taking Back Sunday
Artist 2: Evanescence
Artist 3: Elliott Smith
Artist 4: blink-182
```

In [8]:
```
top_five=bestModel.recommendProducts(1059637,5)
key=artistData.keys()
for i in range(0,5):
    a=top_five[i][1]
    if a in key:
        print "Artist "+ str(i)+ ": "+ str(artistData.get(a))
```

```
Artist 0: blink-182
Artist 1: Elliott Smith
Artist 2: Taking Back Sunday
Artist 3: Incubus
Artist 4: Death Cab for Cutie
```

In [ ]: