DATA 643 Proj 2

Dan Fanelli

June 12, 2017

Content-Based and Collaborative Filtering

The last.fm story continued...

Figure 1:

Listen counts from Proj 1:

Below is a sample of the initial join between ***users*** and ***artists***, along with the combination's corresponding ***listen_count***.

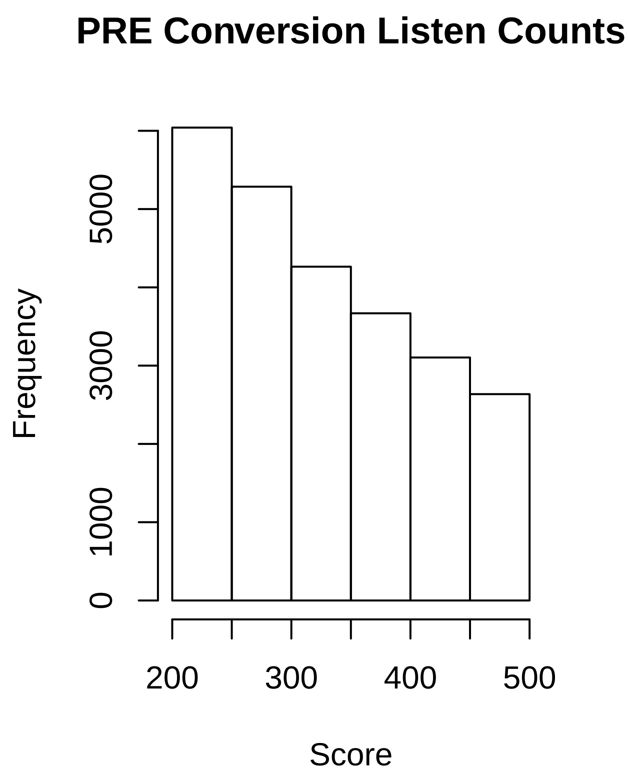
Table 1: A Sample of the Initial Data (Similar to Proj 1)

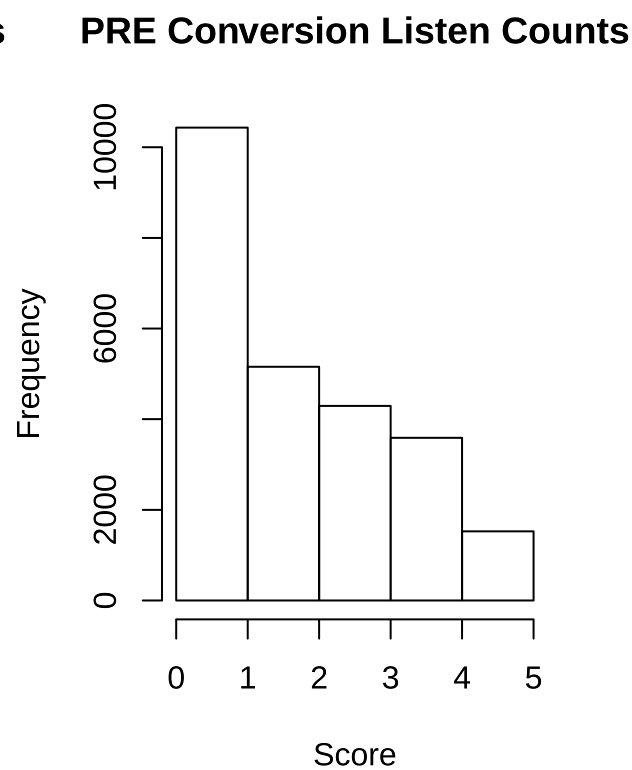
	userID	artistID	artist	listen_count
47647	1158	813	As I Lay Dying	242
38642	953	546	The Ting Tings	331
36731	1337	3255	Athlete	355
47449	117	227	The Beatles	243
30878	1160	12792	Sp7	440
40895	2042	4177	Switchfoot	306
45047	827	952	Skid Row	264

	userID	artistID	artist	listen_count
50602	1816	2133	Milburn	218
43233	1989	18035	Gary B	282
31625	1233	2018	September	428

Listen counts as SCORES from 0 to 5:

The listen counts above are normalized to a SCORE of 0 to 5. The pre-normalization and post-normalization histograms are displayed.





Listen counts as SCORES from 0 to 5:

Below is a sample of the post-normalization scores.

Table 2: The Core Listen Count Data Simplified into Rankings from 1-5

	userID	artistID	artist	listen_count
47647	1158	813	As I Lay Dying	1
38642	953	546	The Ting Tings	2
36731	1337	3255	Athlete	3
47449	117	227	The Beatles	1
30878	1160	12792	Sp7	4

	userID	artistID	artist	listen_count
40895	2042	4177	Switchfoot	2
45047	827	952	Skid Row	1
50602	1816	2133	Milburn	0
43233	1989	18035	Gary B	1
31625	1233	2018	September	4

A Sample of the User-Artist Matrix:

Only a sample since the x-axis corresponds to all ***7251*** artists and the y-axis corresponds to all ***1588*** users.

userID	a1	a2	a3	a5	a6	a7	a8	a9	a10
127	NA	NA	NA	NA	NA	4	NA	NA	NA
135	NA	NA	NA	NA	NA	NA	NA	5	NA
139	NA	NA	NA	NA	NA	3	NA	NA	NA
172	NA	NA	NA	NA	NA	1	NA	NA	NA
179	NA	NA	NA	NA	NA	1	NA	NA	NA
213	NA	NA	NA	NA	NA	4	NA	NA	NA

User-User Filtering ("UBCF") Recommendations

Below are the ***"Top 8 User-User"*** Recommendations for specified users.

Table 4: UBCF: User-User topNList

User	Rec_1	Rec_2	Rec_3	Rec_4	Rec_5	Rec_6	Rec_7
1	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	DIR EN GREY	Combichr
2	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	Marilyn Manson	DIR EN
3	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	DIR EN GREY	Combichr
4	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	DIR EN GREY	Combichr
5	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	Marilyn Manson	DIR EN
6	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	DIR EN GREY	Combichr
7	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	Marilyn Manson	DIR EN
8	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	Marilyn Manson	DIR EN
9	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	DIR EN GREY	Combichr
10	MALICE MIZER	Diary of Dreams	Carpathian Forest	Bella Morte	Moonspell	DIR EN GREY	Combichr

Item-Item Filtering ("IBCF") Recommendations

Below are the ***"Top 8 Item-Item"*** Recommendations for specified users.

```
## Available parameter (with default values):
## k     =  30
## method    =  Cosine
## normalize     =  center
## normalize_sim_matrix  =  FALSE
## alpha     =  0.5
## na_as_zero    =  FALSE
## verbose   =  FALSE
```

Table

User	Rec_1	Rec_2	Rec_3	Rec_4	Rec_5	R
1	Tilly and the Wall	Jag Panzer	Heathen	Sacred Reich	Bloodbound	M
2	Maxwell	Jamal	Absurd Minds	Flaw	Camila Moreno	F
3	Carpathian Forest	DIR EN GREY	Covenant	Pleq	Keane	M
4	MALICE MIZER	Carpathian Forest	DIR EN GREY	Psyclon Nine	Covenant	I
5	MALICE MIZER	Carpathian Forest	Bella Morte	DIR EN GREY	Psyclon Nine	C
6	Bella Morte	The Beatles	Dilated Peoples	Monica	Britney Spears	T
7	INXS	Radiohead	Green Day	Racionais MC's	Christina Aguilera	F
8	Psyclon Nine	Gothminister	Sparklehorse	Black Eyed Peas	Nelly Furtado	A
9	VAST	Buena Vista Social Club	People Under the Stairs	Cine	Page France	A
10	Dawn of Ashes	Kylie Minogue	Marc Almond	INXS	Pleq	T

Conclusions

Ideally, code like below could run on multiple machines, but RAM and Time did not allow:

```
for(a in 1:floor_listen_count_options){
  for(b in 1:ceiling_listen_count_options){
    for(c in 1:number_of_ratings_blocks_max){
      for(d in 1:num_nearest_neighbor_options)
        #etc. etc.
        do_the_calculations(data, a, b, c, d, ...);
    }
  }
}
```

This problem shows how useful a technology like Spark could be in distribution all of these possible combinations to the RAM of multiple computers simultaneously.

When using 5000 as a sample size, all recommendations came back very very similar. With 25k, they got a bit more unique

The run times for learning sample sizes of:

- 5k = 4 minues
- 25k = 40 minutes
- All = (not happening)

Beyond this, the main thoughts were that the User-User Filtering seemd to produce more duplicates than the Item-Item Filtering. There didn't seem to be an obvious reason why certain bands were showing up the most often - ie - MALICE MIZER was first on nearly all the lists, it was not because this artist appeared more often than the others, or any other obvious reason.

Final Thought: Watching the video about Spotify and how they need to use a subset of the full matrix is hitting home as my 25k run goes into minute 35.