RJ Duran
MAT259 Winter 2012
Data Visualization
Project 3

Introduction

The goal of this project is to correlate two sets of data and visualize their relationship. The question is “How do last.fm album plays compare to SPL CD checkouts for popular music in 2011?”

In the visualization there are two colors used to represent the percentage of listens on last.fm vs number of checkouts from the Seattle Public Library in 2011. As you hover over each album the colors become visible.

Queries

SPL QUERY

```
select title, count(*), subj  from inraw where year(cout) = '2011' AND
itemtype = 'accd' AND (subj like '%popular%' and subj like '%music%') group
by title order by count(*) DESC;
```

LAST.FM QUERIES

album.search

```
http://ws.audioscrobbler.com/2.0/?
method=album.search&album=fame+monster&api_key=d023a190a026febbe19f38ba8ab53c
58&format=json
```

album.getinfo

```
http://ws.audioscrobbler.com/2.0/?
method=album.getinfo&api_key=d023a190a026febbe19f38ba8ab53c58&artist=Lady+Gag
a&album=The+Fame+Monster&format=json
```

Explanation

SPL QUERY

This query pulls the data for title and total number of checkouts in 2011 for accd itemtypes with a subject like popular and music. The data is then grouped by title and ordered from most to least number of checkouts.

LAST.FM QUERIES

There are two queries used to request the information from last.fm, album.search and album.getinfo. album.search accepts an album title and returns the top results. In this case it’s requesting fame+monster. It’s using an api key and requesting a return data format of json. Once the album is found we know the artist and the correct album title. album.getinfo is used to get additional information like the playcount and artwork url for the album.

Query Request Time

The SPL QUERY doesn’t take more than a few seconds at most but the LAST.FM QUERIES are incredibly slow when doing a search and pulling a lot of data. I decided to save both data sets in text files to reduce the data load time. When the program loads, the data loads immediately. The image files used to display the album artwork still needs to load so this takes a few seconds.

Analysis

I tried to communicate the difference between last.fm album plays and SPL checkouts and ran into a few issues. The biggest issue is that the actual data shows that last.fm is by far the most used method for listening to music. Because of this, the scale for last.fm is between hundreds and millions of plays per album vs zero to a little over one thousand for SPL checkouts. I had to normalize the data and represent it as a percentage to illustrate the difference between the two.

Another big issue has to do with the naming convention for library data. There are anomalies in the title’s used and no artist names are ever used. This means that when the data is passed into last.fm, it has to work hard to auto correct and search for albums.

The correlation between the two data sets also presented a challenge in accurately finding the correct album from last.fm. You will notice that some of the albums have no artwork or stats for playcount. This is because last.fm couldn't find the album in it's records. They rely on outside ID3 tagging for their data.