CMPS 263 Final Project Report

Namrata Simha
npsimha@ucsc.edu

Madhu Shivashankaraiah
mshivash@ucsc.edu

Fa Chang
fachang@ucsc.edu

March 20, 2017

Visualizing Bollywood music from one's favourite artist.

Leading question: How to choose a music genre for a list of songs from my favourite Bollywood singer/lyricist/music composer?

In this Project, we have visualized the music in different genres of music in Bollywood and have retrieved different songs from various artists, i.e. music composers (in our current use-case) of the songs in these movies. These songs are from the many types of music that have been around for the past 50 years in the Indian Film Industry.

Data Design and Process:

(i) Design decisions and description of data:
Hindi Geet Mala http://www.hindigeetmala.net/ is the website that we have chosen to work with for our data. It is a website that provides access to a vast database of Bollywood movie details. We have decided to work with this website and have written a python script to extract and obtain a database of movie and song details and work with that data for our project. We have obtained a .csv of the data for about 37000 songs from Hindi Gett Mala. The extraction of this data is explained in the next section.

Other data sources that we tried to use : Cinemalytics API: http://api.cinemalytics.in We tried this dataset, but determined that the data was too scarce as we weren't able to retrieve all genre information despite scraping data from the API as well as Wikipedia and YouTube for the song details. We also obtained only around 5800 songs from the API as opposed to the 37000 songs that we obtained from Hindi Geet Mala.

Using the data available on the Hindi Geet Mala website, we wrote a python script *makeDataset.py*, available to view at our GitHub. This script extracts every page available on the website by the categories available and separates out the following fields for each song for us to use:

- *genres* - Type or genre of song.
- *songNames* - Names of the songs.
- *singers* - The signer who sung that particular song.
- *musicDirectors* - The name(s) of the person/people who composed music for the song.
- *lyricists* - The name(s) of the person/people who composed lyrics for the song.

- *movies* - The movie that the song is from.
- *years* - The year in which the song was released.

This extraction was done from the HTML scripts of the website, such in the following link: Categories page example HTML (Please allow your Adobe Acrobat to open this file.)

(ii) Data Scraping Process:
For reproduction of our code by anybody who has access to our GitHub and this document, here is the simple algorithm with the steps that we followed to write the code that we used to obtain the data for our project:

Step 1: The HTML file is split (using regular expressions[3]) into parts by the category (i.e. the *genre* as we described it, as per the usage on the website) name.

Step 2: For each category found on the currently available 3 pages of categories, do the following:

Step 2(a): We first find the number of pages of songs that are available. For each of those, do the following:

* In each of those pages, we split (again by regular expressions) the page by songs (items).
* For each item found, a new entry into individual arrays for each of the fields described is added.

Step 3: Write all accumulated song entries into a .csv file[2] called *songs.csv*

Thus we have songs and details for each category which we describe as *genre*. This adds up to around 37000 song entries overall which we have written into a .csv file[2] called *songs.csv*.

Python Script for Data Wrangling:

We have written a script *makeDataset.py* which is described in the previous section to extract the dataset that we require for this project, available to view at our GitHub:
https://github.com/namrata-simha/Bollywood-Music-Visualization

Input - Output:
The **input** to the *makeDataset.py* is the HTML source pages of the "Hindi Geet Mala" website as described in the data design section.

The **output** from the *makeDataset.py* is a .csv file containing the fields described in the data design section. These fields each have around 37000 entries, as mention before. This .csv is available on our GitHub repository. The file is called *songs.csv*.

Interactive Visualization using a Java Desktop Application:

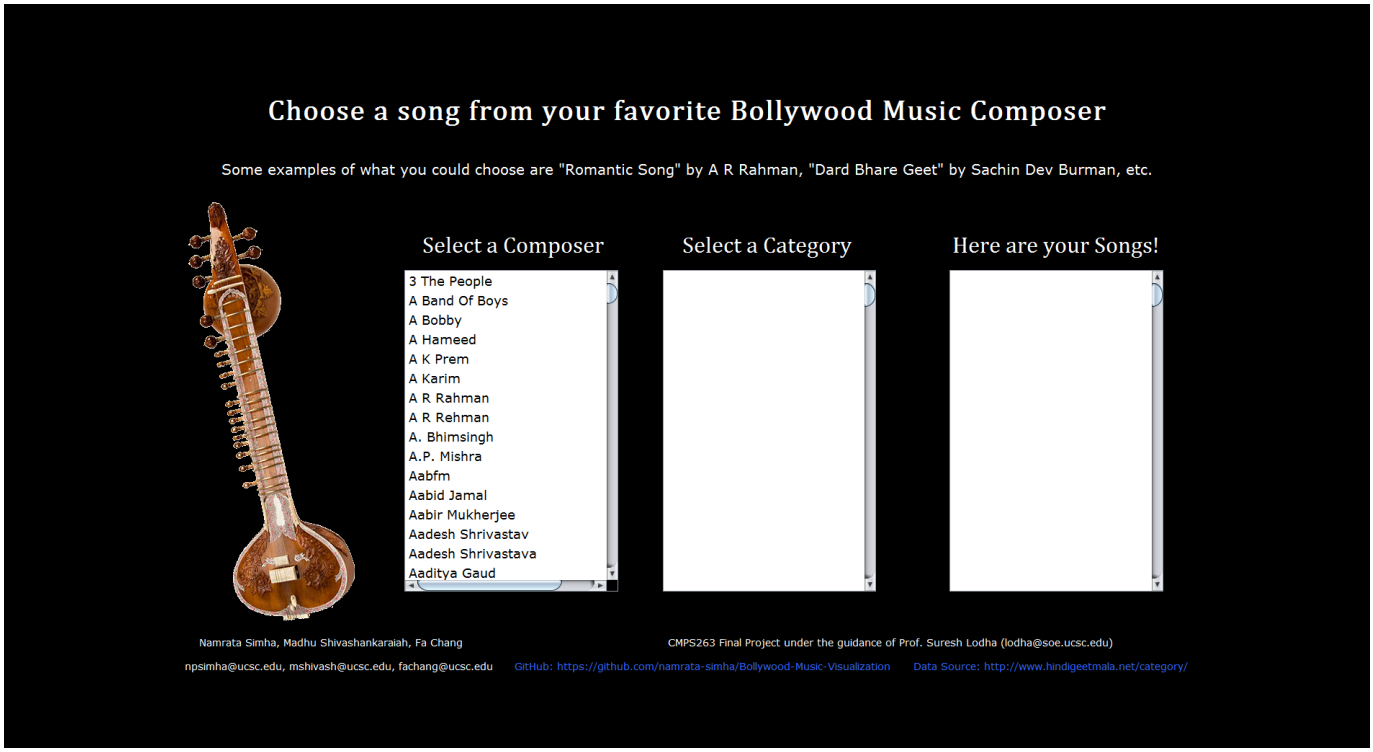
After extracting the data in the above mentioned format, we built a Java application using the GUI NetBeans[10]. Since most of us didn't have a full working knowledge of NetBeans and Java desktop applications, we referred to various tutorials among which two were particularly helpful, and hence we based on work on these tutorials[11][12].

The application that we built takes the data (*songs.csv*) and presents it in 3 windows that allows a user to select a (i)*Music composer* of their choice, then a corresponding (ii)*Category* of song by that composer and then finally a (iii)*Song* that is of the selected category (or *genre* - genre and category are used in this paper and in the code interchangeably) by the particular music composer. This in turn opens a search of the song on the popular bollywood music website https://www.saavn.com.

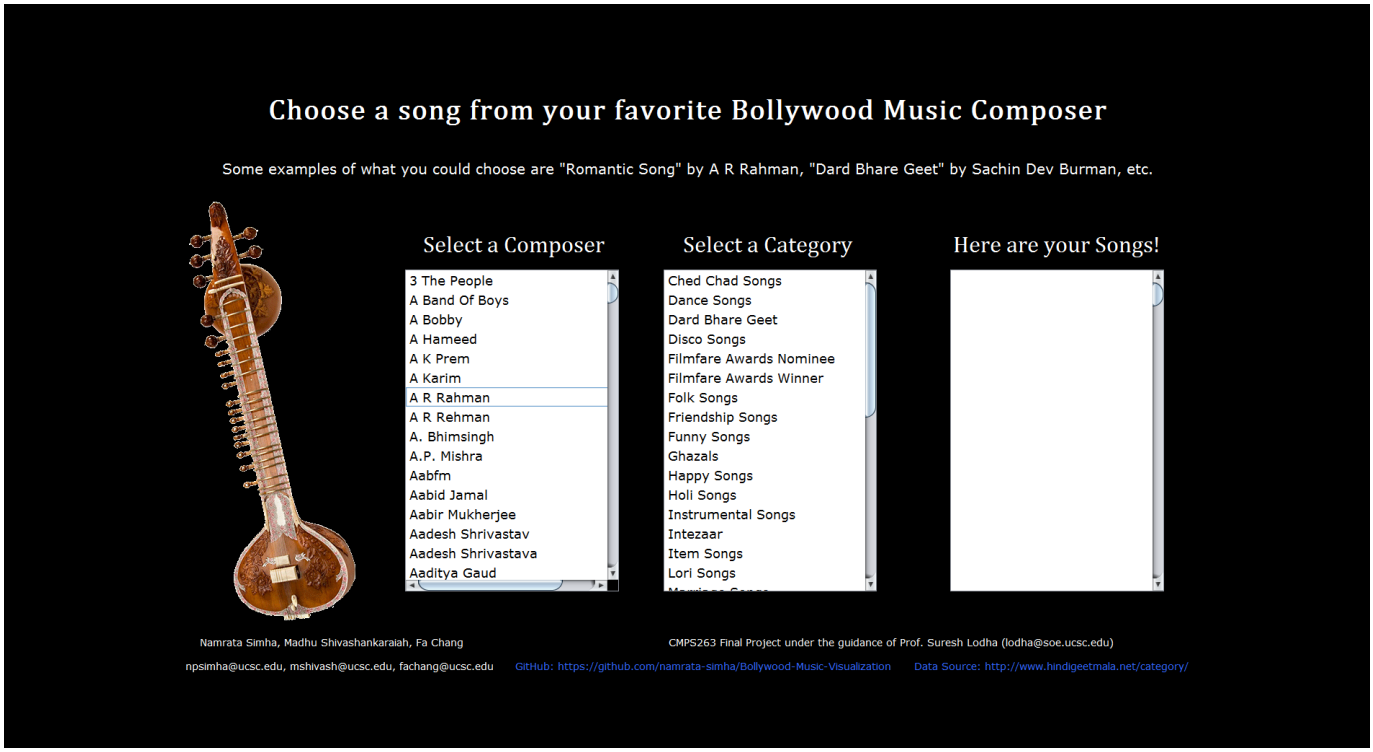
This application can be viewed double-clicking the file *Submission/A.Visualization/BollywoodViz.jar* or clicking on this link: BollywoodViz.jar(Clicking this link may not be allowed in many systems due to security issues. If so, please follow the path aforementioned)

This visualization is interactive and is intuitive to understand and play with. It asks the user the question "Choose a song from your favourite Bollywood Music Composer" and mentions examples of how to select from the 3 fields provided. To help in understanding the layout of the application, here are some screenshots of the working of one full view of the application:

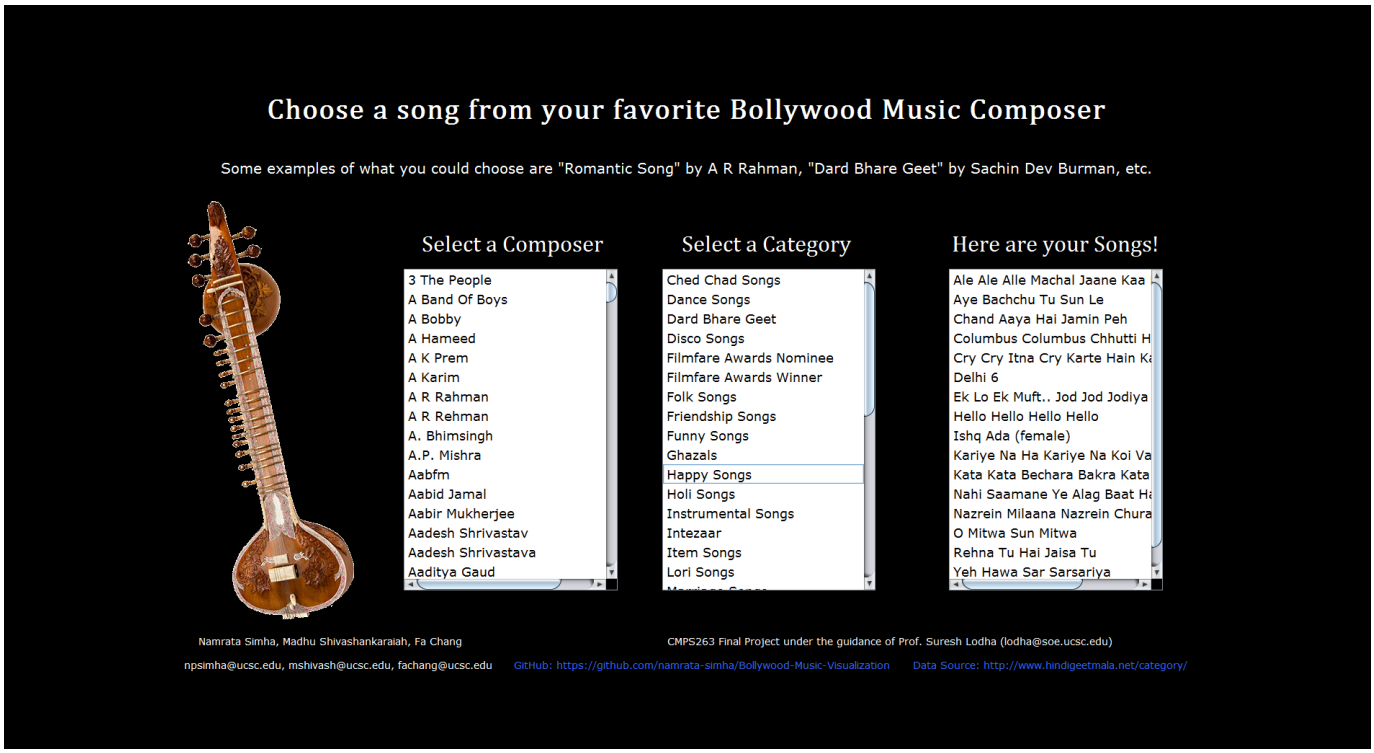
First select a music composer list given:



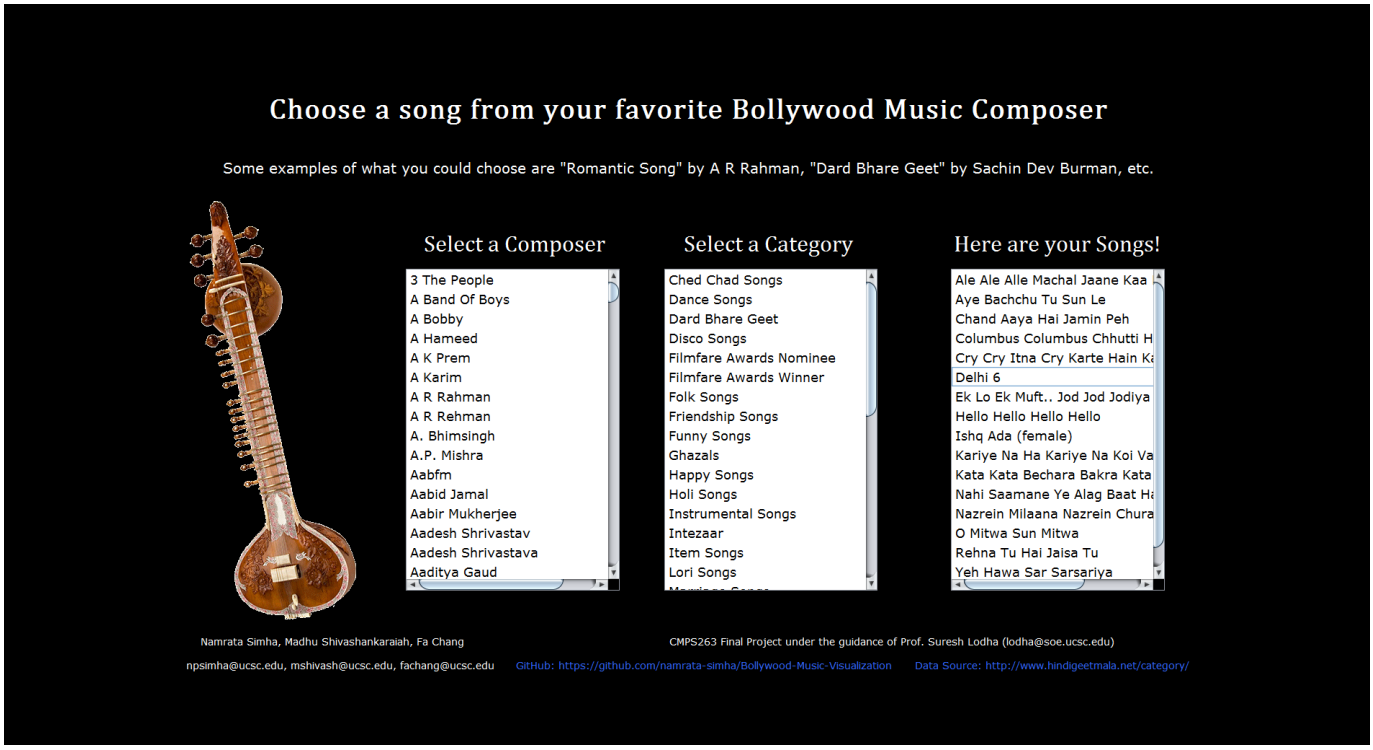
For the selected music composer, a genre/category list is generated. Select one of those:



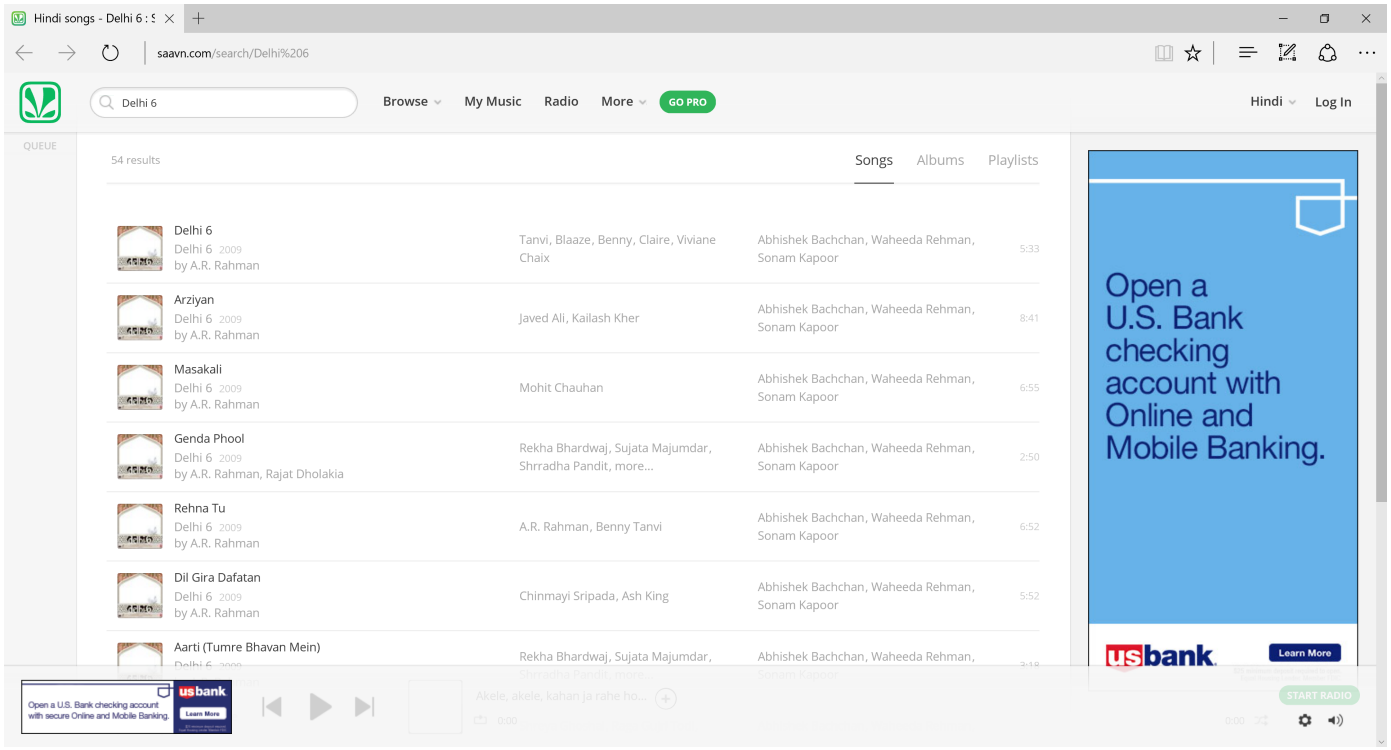
For the selected music composer and genre/category, a song list is generated. Select one of those:



Once a song is selected, as shown:



The song is searched on www.saavn.com and you can listen to it now:



Conclusion:

In this project, we built an application that takes the data scraped from http://www.hindigeetmala.net/ as input and visualizes the selection and search of our favourite Bollywood songs based on the Music composer and category of song using a Java desktop application.
In the future, we can extend this project to include searching by more attributes of the song, such as singer/lyricist/movie/actors etc. Further, there is a lot of misspelling of names and songs in the the dataset that we have collected which can only be cleaned manually by someone with a good exposure to Bollywood. Since we were aiming to finish this project within the quarter, we didn't find the time to look into this matter. Also, for a wider audience, as per a suggestion in class, we could translate some of the categories available and streamline them a little further to make it understandable to non-Hindi speakers.

References

[1] Python documentation for urllib, available at https://docs.python.org/2/library/urllib.html.

[2] Python documentation for csv, available at https://docs.python.org/2/library/csv.html.

[3] Python documentation for re (regex), available at https://docs.python.org/2/library/re.html.

[4] Python documentation for itertools, available at https://docs.python.org/2/library/itertools.html.

[5] Million Songs Dataset:
https://www.opensciencedatacloud.org/publicdata/million-song-dataset/

[6] Million Songs Dataset - Python code to access it:
http://labrosa.ee.columbia.edu/millionsong/pages/code#python

[7] Worldwide music evolution of genres from Google: https://research.google.com/bigpicture/music/

[8] Silicon Valley Data Science project on visualizing the history of 100 years of rock music:
http://www.svds.com/rockandroll/#thebeatles

[9] FlowingData collection of music based visualizations: https://flowingdata.com/tag/music/

[10] NetBeans available to download at: https://netbeans.org/downloads/

[11] NetBeans tutorials available at: https://netbeans.org/kb/docs/java/quickstart.html

[12] Desktop application building using NetBeans tutorial available at:
http://www.javaguicodexample.com/javadesktopguimysql1.html