**CS838 Project Stage 4 Report**  
Fangzhou Mu, Nafisah Islam, Meera George

1. **Datasets**

We combined samples of *tracks.csv* and *songs.csv*. The samples, *track\_sample.csv* and *songs\_sample.csv*, contain ***53935*** and ***100000*** tuples, respectively. ***32000*** tuple pairs survived after blocking and were stored in *pairs\_passed.csv*. The matcher developed in Stage 3 was applied to the candidate pairs. The set of matches is stored in *matches.csv*. A final dataset, *merged\_data.csv*, was created by merging matched tuples.

All datasets are available at [Github Link](https://github.com/nafisahis/CS838-Data-Science-Project/tree/master/stage-4/dataset).

1. **Data merging**

We conducted the following three steps.

1. The matcher was applied to the candidate pairs and a set of matches was obtained.
2. With respect to each match, the IDs were used to locate associated attribute values (i.e. *song\_title, year, artists*) in both tables.
3. The following rules were used when we created the final table.
4. For *song\_title* and *artists*, we selected ***the longer string*** from the corresponding two tuples, assuming that it is more comprehensive.
5. For *year*, we selected ***the smaller value*** from the corresponding two tuples, assuming that it is more likely to be when the song was composed.
6. When an attribute value is missing in one of the two matched tuples, we put the only value available (from the other tuple) in the final table.
7. In case that an attribute value is missing in both tables, we left the value in the final table blank.
8. Unique attributes in either table were simply carried over to the new table.
9. **Statistics**

The schema of the final table is ***Songs-Tracks****[movie\_title,year,episode,song\_title,artists]*. We may combine this table with *movies.csv* in Stage 5 to obtain another table for data analysis. The table contains ***7280*** tuples. Some sample tuples in the final dataset are shown below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **movie\_title** | **year** | **episode** | **Song\_title** | **artists** |
| one tree hill | 2007 | running to stand still (#5.10) | only fooling myself | kate voegele |
| t in the park | 2015 | weekend highlights: part 4 (#1.64) | what became of the likely lads | the libertines |
| beavis and butt-head | 1994 | date bait (#4.18) | god of emptiness (from the album covenant) | morbid angel |
| la virgen de los sicarios | 2000 | NaN | el santo cachon | romualdo brito+los embajadores vallenatos with robinson damian |

1. **Code**

Code for data preprocessing, blocking and matcher development was completed in Stage 3. Code for data merging, *Merging.ipynb*, is available at [Data Merging Link](https://github.com/nafisahis/CS838-Data-Science-Project/blob/master/stage-4/Data%20Integration.ipynb).

