Project Stage V

Data Analysis

Team

Anna Chang
Sripradha Karkala
Simmi Pateriya

Statistics on Table E (Merged Table)

The merged table (Table E) contains 1331 tuples. Below is the schema and sample tuples:

song_id	track_id	title	artists	song	year	episode	episode_year
457	140954	Lockie Leonard	angus stone+julia stone+angus	Mango Tree	2006	New and Improved (#2.1)	2007
1203	444229	Frank Sinatra: A Man and His Music + Ella + Jobim	paul mann+stefan weiÃŒÃ™+frank sinatra	Put Your Dreams Away	1958		1967
1935	121170	Intervention	scott klass+the davenports	Five Steps	2000	Adam (#9.2)	2005
1935	121498	Intervention	scott klass+the davenports	Five Steps	2000	Salina and Troy (#2.5)	2005
4202	541635	Ministry: Tapes of Wrath	ministry	The Land of Rape & Honey (Live)	0		2000
4432	208737	Shameless	the high strung	The Luck You Got	2005	The Sins of My Caretaker (#3.5)	2011

Data Analysis

Task 1: Discovering anomalies and outliers

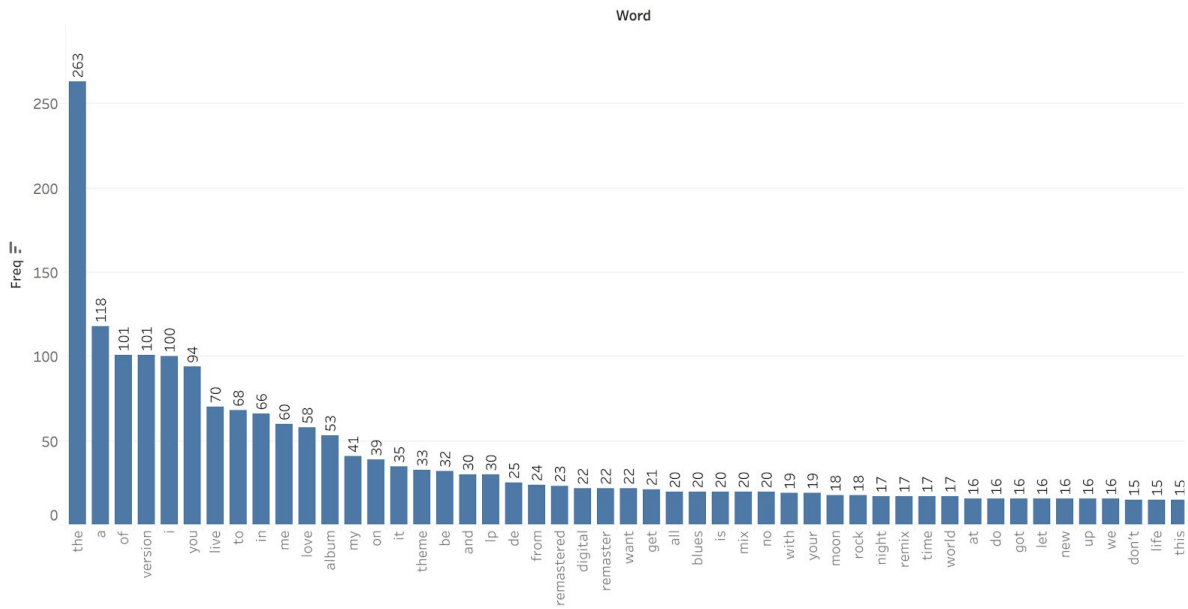
We wrote a script that detected if the year a song was released was later than the year of the TV episodes. This would suggest that a TV episode somehow used a song before the song had even been created. There were 247 such tuples or roughly 18.5% of Table E, and

the full list of anomalies can be found here. After further investigation, we came up with the following explanations for these discrepancies:

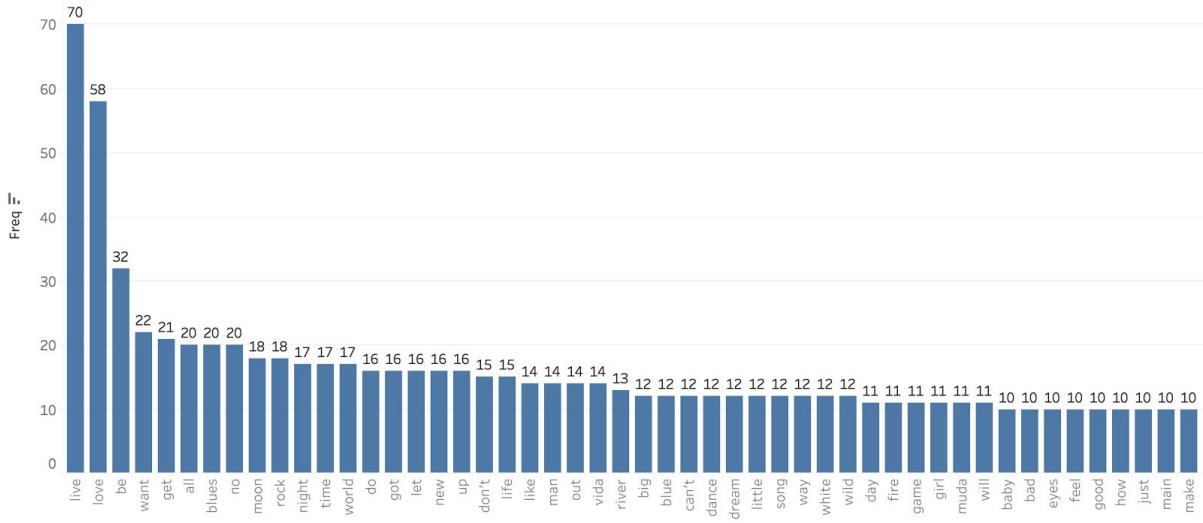
1. Sometimes it was the case that the Song table represented a cover, remix, live, or other version of a song listed in the Track table. This means that the TV episode must have used an earlier version of a song than the one indicated in Table E.
2. In some cases, the original song was an instrumental and was released earlier than the years mentioned in the both tables. We were able to discover this by looking for the original singer which is included in the list of artists.
3. It's also possible that some songs were created specifically for a TV show and later recorded or added to an album.

Task 2: OLAP

We wanted to gain some insight into the subject matter of the songs so we did an OLAP style query by tokenizing the song titles by word and then aggregating them by frequency. Our initial histogram (see below), showed that the most common words were stop words (prepositions, pronouns, articles) and song metadata such as "version", "album", "remastered", "theme", and "lp."



After removing these "stop" words from the histogram, we found that "live" occurs the most frequently in the titles, followed by the word "love."



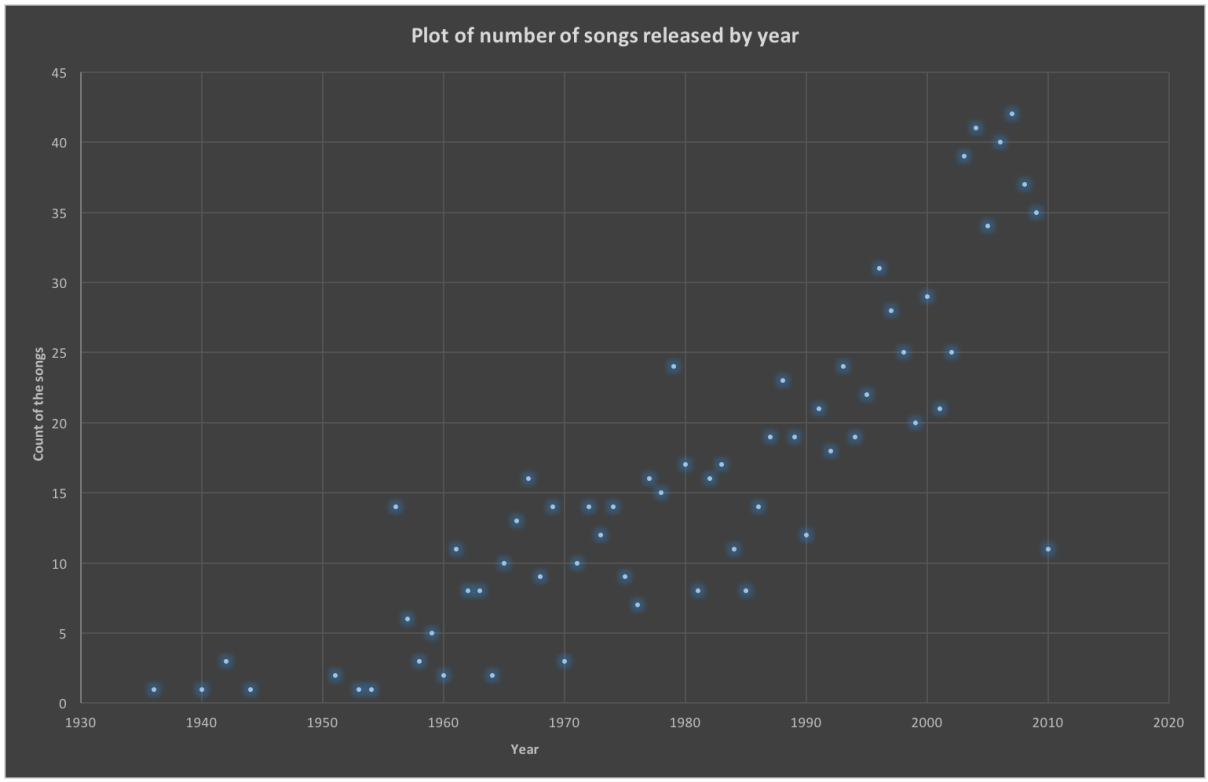
This was surprising since we thought most songs would be about love, so we drilled down by looking only at song titles with the word “live.” By doing so, we discovered that “live” is also a stop word in the song context and 66 out of 70 song titles were using it to describe a “live performance.” Here are some sample tuples:

song_id	track_id	title	artists	song	year	episode	episode_year
11931	302892	Top of the Pops	spice girls+jon b.+eliot kennedy	Say You'll Be There (Live)	1998	(1996-12-25)	1964
21127	44578	Charmed	sarah mclachlan	Building A Mystery (Live)	1999	Something Wicca This Way Comes (#1.1)	1998
35758	643563	The 70s	barry gibb+robin gibb+maurice gibb+the bee gees	You Should Be Dancing [Live - Las Vegas 1997] (Album Version)	1998		2000
37071	402831	D2: The Mighty Ducks	Queen + Paul Rodgers	We Will Rock You (Live In Ukraine)	0		1994

Therefore, “love” did have the highest frequency (58), although 29 song titles did mention “life” or “vida” (which means life in a number of languages). The 14 instances of “vida” all came from the same portugese soap opera, “Bem-Vindos a Beirais.” We also drilled down on “moon” and found the song “Moon River” was used in 7 distinct TV shows between the years 1962 and 2001, demonstrating an impressive longevity.

Task 3: Correlation Discovery

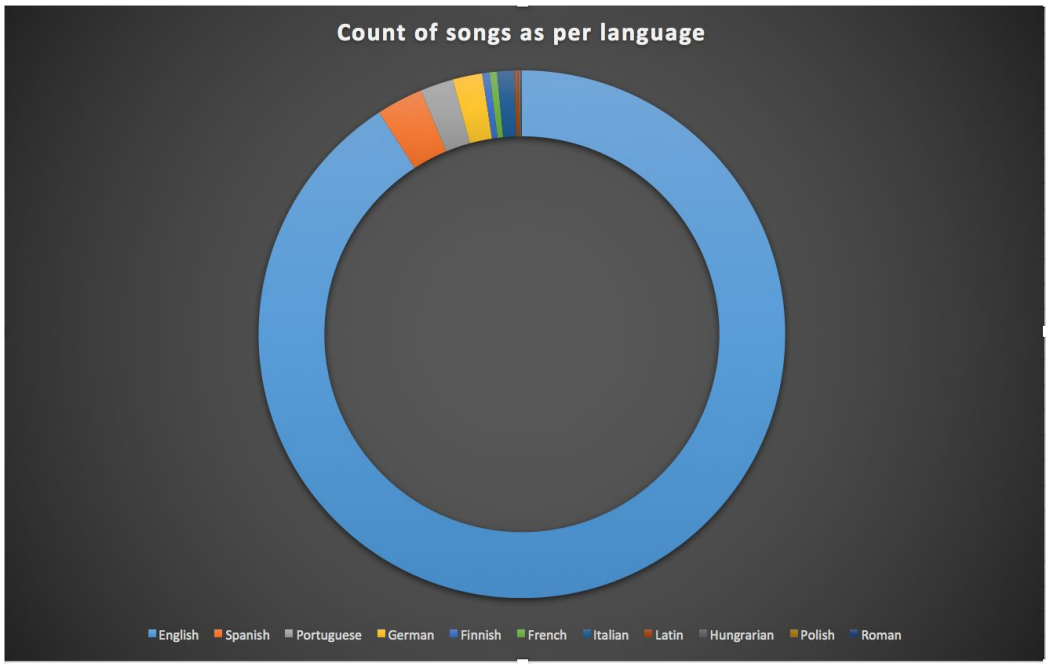
We wanted to see if there was any correlation between the number of songs and the year they were released. The scatterplot indicated that there was indeed a strong positive correlation:



Task 4: Classification

We also decided to experiment with the songs and detect the language based on their names. It is not surprising that most of the songs (91%) are in English. We found that the next major language in the dataset was Spanish (3%) closely competing with Portuguese and German (2%) each.

This was executed using *langdetect* from Python. After cleaning up, the results are as shown below



Accuracy Numbers

We did not use any Machine learning models for classification or clustering in this stage. However, the classification of songs based on languages using *langdetect* package produced an accuracy of 64% . This could be due to the fact that songs names are usually small (4-5 words) and these words could span across similar languages leading to incorrect prediction.

Conclusion & Next Steps

Our anomaly detection showed that we might not get true matches using the data from the Song and Tracks tables. In other words, even if the song titles match, the song may have been sung by a different artist in the TV episode than the one we matched to. To improve the accuracy of matches, we can change our matcher in Stage III to match not only on artist name and song title, but also on year (i.e. Track Year must be later or equal to Song Year).

Our OLAP queries confirmed that most songs really are about love (at least in our dataset) and also suggested room for improving our Stage III matcher. Specifically, it helped to identify stop words that we were unaware of, such as “live.”

One issue we encountered in this stage was that we downsampled during Stage III and so Table E may not reflect the true distribution of the whole dataset. Moreover, even if we didn’t downsample, we don’t know how the original data was acquired and this can affect how we interpret our results.

Given more time, we’d like to delve deeper into some of our results. In particular, we saw a positive correlation between the number of songs and the year the song was released.

Would this still be true if we were to look at all the data in Songs? Depending on how the data in Songs was acquired, this can tell us about the music industry and perhaps reflect an increase in the number of songs recorded each year.

Overall, our analysis found some interesting patterns in our data and this analysis can be strengthened by looking at all the data in the Song and Tracks tables rather than just a small subset and by gaining more insight into the provenance of the datasets.