

Steven universe lyrics

Big Data Project



Chauvet tIMOTHE – mathieu camille

Problem definition:

Steven universe is a show with a plethora of musical segment, especially songs accompanying character arc development, representing the growth of each of their respective journey. As each character battle with different issues, there are recurring thematic coming up through their songs, as such we would like to be able to identify the character singing by the lyrics of the song.

The main challenge is that even through the approach are different the song often revolve around very similar themes, “love”, ”strength” … We might be able to simply count the occurrence of a certain type of vocabulary and attribute it to a specific character, but it may not be sufficient. We will also have to proceed with further analysis and see the interaction between the words.

# Methodology:

First, we created the database by fetching the song title and lyrics used in the show. To do so we used a web scrapper: lyricsgenius specially design for the website genius who compile songs lyrics and classify them.

To start we must connect to an API, the genius programming interface to be able to use it, then we just had to scrap the ones whose artist was stated to be “Steven Universe”, meaning they appeared in the show. The challenge here is to specify which characteristics of each elements we want or not, for example we are only interested in the song’s title, its lyrics and the singer, not the album’s title or the type of section (chorus, etc.) or purely musical pieces.

## Searching the lyrics and who’s singing them

Python, like many other languages, has a RegEx library that allows us fetching certain text patterns. This functionality will be helpful in our project as the lyrics from Genius are messy. Therefore, we did these patterns:

* **Finding who’s singing**, which is quite simple as the result data is written like a theater script
* **Finding the lyrics**, much more difficult as we had to find patterns between two colons and remove the last word (corresponding to the next person singing)
* **Removing annotations** that are between parenthesis in the lyrics

After that cleaning process, we were able to quickly remove unwanted songs that doesn’t have multiple characters singing, as these songs can’t be automatized since the character singing the song isn’t written on them.

## Exporting the data

Then we need to convert the scrapped songs to a format usable in the database, we choose Json since it enables us to keep the formatting of the data.

We decided to use mongoDB since we had trouble with the virtual machine. More precisely we used pymongo, a library that allow to use mongoDB through python. Seen as we’ve run into some trouble to start the mongoDB server it might be useful to precise that first the task manager need to be openend to make sure that the mongoDB server is running. Then the command “mongod” and “mongo” need to be entered in a command windows.

Then we can directly connect to the server and create the database along with it’s component inside our python program.

## About the data size

We made the whole process using only a limited number of songs to save on time and see if the process was viable. On the genius website there are more than 20 albums corresponding to the singer “Steven Universe”, each with up to 40 titles, therefor we have to work with an initial dataset of up to 1000 elements. To reduce this number, we proceeded to a pre-selection: we scrapped only the pieces having lyrics. Then we eliminated duplicates and hand-searched if there were title who did not appear in the show or who were not sung by characters of the show.

# Results and discussion:

# Guideline: