

WORD CLOUD WITH SHINY

Ufuk Cem BIRBIRI

April 10, 2022

1 Word Cloud

1.1 Users Description

Users are people interested in lyrics of songs and genres. Lyrics are words used in songs. Users can visualize what kind of lyrics and how many of them are used in a specific genre with a word cloud. To do that, I used lyric summaries of the songs because complete lyrics are not publicly available, however lyric summaries of the songs are available. So as a summary, I grouped the songs by genre and took their lyrics summaries then visualize them with a word cloud. More detailed explanation is in the rest of the document.

1.2 Visual tasks and the visualization goals.

The visual tasks in Word Cloud are shown in the table below.

User task	Details
Explanation of word cloud	Give info about what is word cloud
Explanation of stopwords	Give info about what is stopwords
Explanation of the app	Give info about app and wasabi dataset
Choose a genre	Choose a genre to display its lyric
Remove stopwords	Removing the stopwords from the lyrics
Remove specific words	Removing words given by user
Show the count of the word	When user hoover on words on word cloud, show the count
Select years	Select a range of years to display lyrics in that range
Show lyrics that don't have publication date	Show lyrics that publication date is NA
Show count of songs	Display the number of songs in the given years
Choose word cloud theme	Select the word cloud visualization theme(Dark, Barbie,...)
Select shape	Select shape of the word cloud (diamond, circle, ...)
Select size	Select size of the word cloud

What are stopwords?

Stop words are basically a set of commonly used words in any language, not just in English. The reason why stop words are critical to many applications is that, if we remove the words that are very commonly used in a given language, we can focus on the important words instead.

Some examples of English stopwords:

i, me, my, myself, we, our, ours, ourselves, you, your, yours, they, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, do, does, did, a, an, the, and, but, if, or, because, as, until, s, t, can, will, just, don, should, now, ...

1.3 Attributes from the WASABI dataset

The following attributes will be needed from the song field:

- Song field
 - `_id` (Song id)
 - `genre` (type:string)
 - `summary` (Summary of lyrics in a few lines, type:string)
 - `publicationDate`

1.4 Processing raw data

Data manipulation is done in the following order.

- Load and analyze data.
- Clustering the genres.
- Assign each song to a cluster of genre.
- Group songs by genre, clear lyrics, create csv files for each genre.

You can find the detailed explanation of these steps below.

1.4.1 Load and analyse the data

The data is loaded from wasabi web page and useful columns are selected such as song id, genre and publication date. The empty and not-given genres are detected and replaced by "Not specified". Same method is used for publication dates. The empty and not-given publication dates are detected and replaced by "0000".

Given publication dates are cropped and only the year is taken, month and day information is ignored. Example: 1998-06-22 becomes 1998.

In the wasabi dataset, one song can belong to multiple genres. In this situation, the genres are splitted. Example:

- `ObjectId(5714dec325ac0d8aee38093d)`, "metal, punk, rock"

becomes:

- `ObjectId(5714dec325ac0d8aee38093d)`, "metal"
- `ObjectId(5714dec325ac0d8aee38093d)`, "punk"
- `ObjectId(5714dec325ac0d8aee38093d)`, "rock"

where the `ObjectId(...)` is the song id.

1.4.2 Clustering the genres

There are many genres in wasabi datasets. Some genres are originated from other genres by years. For example, 'Skiffle' is originated from genre 'Folk', or 'Emo' is originated from 'Punk'. At this step, different genres are clustered according to their ancestors. For example, samba, bachata and salsa are grouped under the Latin music. In addition the Skiffle music genre is originated from folk music so it is grouped by folk music genre. Below, you can find the main genre and its cluster members. It is good to remember that a new genre can be a mixture of multiple genres. Here,

clustering is made by taking into account the most similar genre.

Clusters of genres				
Rock	Pop	Metal	Hip-Hop	Blues
-Rock music -Surf music -Industrial music -British Invasion -Shoegazing -Minneapolis sound -Neue-Deutsche Welle -Experimental music -Beat -Dark cabaret -Palm Desert Scene -Adult album alternative -Motown	-Pop music -Wall of Sound -Beach music -Vocal music -Music Hall	-Metal music -Deathcore -Deathgrind	-Hip-hop music -New jack swing -Miami bass -Jumpstyle -Hands Up -Hyphy -Lo-fi -Crunk -East Coast hip-hop -Swing revival -Breakbeat -Yé-yé -Police procedural -Snap	-Blues music -Liedermacher -Boogie-woogie -Bolero -Tulsa Sound Singer-songwriter
Dance	Electronic	Jazz	Raggae	Country
Dance music Disco Hi-NRG Garage Speed garage Freestyle Bhangra Cabaret Low fidelity Sirtaki Tropicália	Electronic music Big beat Grime Électronique Musique-électronique Remix	Jazz music Quiet storm Afrobeat Dixieland Stride Screamo Tin Pan Alley Cumbia	Raggae Dubstep Reggaestep Hardstep Ska Dub music Oldschool jungle Old-time Mento Traditional black gospel Ragtime	Country music Honky-tonk Music of Lubbock Texas Music of Ireland Western music Moombahton Kuduro

Funk	House	Rap	Folk	Soul	Punk
Funk music Show tune World music Go-go	House music Lullaby Teenage-tragedy song Plunder-phonics Balada Gaana	Rap music Ballad Freestyle music	Folk music Skiffle Tejano Baggy Topical Music of Scotland Boogie Ballet	Soul music Stoner Kwaito Highlife	Punk music Emo Ragga PBR Mariachi Candombe Junkanoo
New Wave	Psychedelic	Christmas	Trance	Techno	Grunge
New wave Cumbia New-age music	Psychedelic music	Christmas music	Trance music Contemporain	Techno Downtempo	Grunge Post-grunge
Latin	Contemporary	Classical	Hardcore	Harmony	
Latin music Bachata Samba Salsa Bossa nove Pasodoble Rumba Flamenco Zumba.	Contemporary Madchester Baroque	Classical music Orchestra Crossover Ambient Patriotic Piano Waltz Música sertaneja	Hardcore music California Sound Exotica	Harmony music Doo-wop Drone Worldbeat Schlager A cappella Music of Italy	
Instrumental	Poetry	Acoustic	Religious	Celtic	Comedy
Instrumental music Minimal Soundtrack March Musique concrète Kayōkyoku Jimmy Buffett Circus Calypso Doctor Who fandom	Poetry Ballade Chanson	Acoustic music Hymn Easy listening Bebop	Religious music Christian Gospel Siren Spoken Feminism	Celtic music	Novelty music Parody music Comedy music

1.4.3 Assign each song to a cluster of genre

With a for loop, every song is assigned to a cluster of genre. A new column is created for this assignment that is called 'grouped_genre'. This new column is used for grouping the songs by genre.

1.4.4 Group songs by genre, clear lyrics, create csv files for each genre.

Songs are grouped by genres. Then, songs are matched with corresponding lyrics. These lyrics are cleared before writing a file. Clearing is removing the punctuation and space. For example:

I will! becomes I will

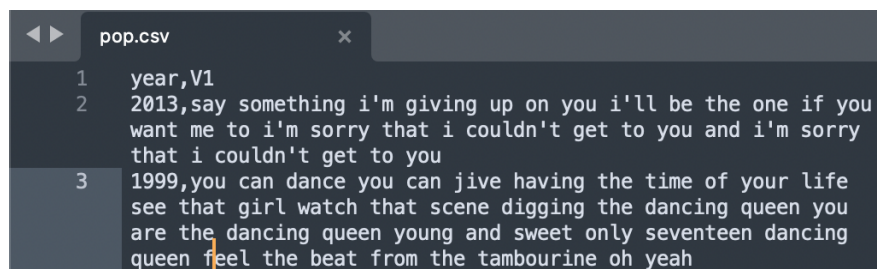
Then, a .csv file is created for each genre which has lyrics and publication date. The data is ready for word cloud in visualization.

1.5 The visualization technique

I will use Word Cloud to represent words in each genre. The size of the word in the tag cloud will represent the frequency of the word i.e if a word is used a lot in lyrics, its size will be bigger than others. Lyrics are made of words so word cloud is a good choice to visualize lyrics to see how often the words are used.

1.6 A visual mapping of variables

Each genre will has its own word cloud since each genre has its own csv file. In this file there are two columns. The first one is the publication year of the song and the second one has the lyrics summaries of the song. The number of rows is the amount of songs belong to that genre. The Fig 1 shows the first three rows of the pop.csv file which has lyrics and publication date of songs for genre pop.



```
pop.csv
1 year,V1
2 2013,say something i'm giving up on you i'll be the one if you
  want me to i'm sorry that i couldn't get to you and i'm sorry
  that i couldn't get to you
3 1999,you can dance you can jive having the time of your life
  see that girl watch that scene digging the dancing queen you
  are the dancing queen young and sweet only seventeen dancing
  queen feel the beat from the tambourine oh yeah
```

Figure 1: The first three rows of pop.csv file. This file includes the lyrics and publication years of the songs. Each row is a different song. The first column shows the publication year and named as 'year'. The second column shows the lyric summaries of the song and named as 'V1'.

Word Cloud is a vizualisation method that displays how frequently words appear in a given body of text, by making the size of each word proportional to its frequency. Colour used on Word Clouds is usually meaningless. I will not use colour, every word will have the same colour. The location of the word is also meaningless. Some words are written horizontally and some are written vertically which are random and do not mean anything. Only meaning feature in the word cloud is the size of the word.

The Fig2 shows the word cloud of genre pop in the Shiny app. The words are visualized proportional to their frequency. Users can see how many times a word is used by passing around the mouse on the words. According to figure, the word 'love' is used 587 times.

Figure 3: Left: User options for song lyrics, publication year, stopwords and the number of words displayed in word cloud. Right: Visualization options for the word cloud.

- Select the number of words in the word cloud (minimum is 3).

Visualization options:

- Choose the word cloud theme. Dark mode is the default one. There are other themes such as Barbie, Jungle, Sunset, and Default(the default theme of wordcloud2 function in wordcloud2 library)
- Select the shape of the word cloud. Options are circle(default), cardioid, diamond, triangle-forward, triangle, pentagon and star.
- Select size of the word cloud from 1 to 2.5 (default is 1.2). When you increase the size, if the words do not fit the page they disappear.
- Remove specific words in the word cloud (maximum 10 specific words).
- Select the number of words in the word cloud (minimum is 3).
- Report a bug or see the code with the given link.

In the Fig.4 the word cloud shows the words of genre 'Religious' with the theme is 'Barbie' between 1984-2017. As you can see, the number of songs is 52. The word 'lord' is used 14 times.

1.7.3 On the main panel:

- The word cloud of the chosen genre is shown.
- When user pass over the words with mouse, the number of words that is used in the lyrics is shown at the left-down corner of the rectangle.
- There is an explanation of the word cloud in the below of the word cloud image. The chosen publication years and number of songs are displayed.

There are two other browsers in the app. "What is world cloud?" explains why word cloud is used in some cases. Also, there is an explanation of stopwords with examples. "About this app" gives information about the Data Visualization course and some explanation of how to use the app. The Fig.5 shows the other tabs.

References