Pop Songs As Mirror of Today's Society and Orbuculum of Future of Music: Textual Analysis of Pop Song Lyrics

Name: Saumyaa Shah NetID: sns9906

Word-Count(With References): 2385

Introduction

In this paper, we are interested in examining trends in pop song lyrics over the years, their significance in human history and their influence on future pop songs and artists.

Pop music, in modern context, emerged as a standalone music genre in the mid-1950's. It originally derives influence from folk music and rock 'n' roll. As delineated in [2], pop music has gone through significant changes in music structure and lyrical content. In early 1960's, structurally, pop music was heavily influenced by blues and rock. In contrast, recent pop music is more towards hip-hop and rap.

Analogous to poems during the Romanticism movement in the 18th century, pop songs can also act as markers in humankind's cultural and political history. As mentioned in [2], socio-economic conditions the artist grew up in, culture and political scenario of society, unforgettable life experiences, etc. have been the primary subjects of early pop songs. For example, many famous artists have composed songs about mental health issues and societal struggles such racial segregation or identifying as queer. Their lyrics not only portray their emotional state of mind, but also reflect the environment they lived in.

Pop music has also inspired young adults to speak up about their issues and express their thoughts through music. However, in recent years, many studies have blamed pop music for destroying creativity among youngsters and promoting obscenity. Additionally, as described in [5], digital pop music is also used as a marketing strategy by enterprises. By creating a song with repetitive lyrics, they aim to control the minds of young adults and gear them towards their services.

In this paper, we perform a textual analysis of lyrics of pop songs from 1950 to 2019. We examine the most common topics in pop songs and most popularly used words for said topics in a time-series fashion. We also study the influence of other genres in pop music and predict the musical style of an artist based on the lyrics of their songs. Finally, we perform a burst-analysis of words over time to study the usage of certain words.

Literature

In 2007, the authors of [6] compiled a corpus of pop song lyrics called Giessen-Bonn Corpus of Popular Music (GBoP) and performed a surface analysis of the vocabulary and grammatical structure of the songs. Various corpus-based analysis such as [8] have been conducted that study the lexicogrammatical structure of lyrics and draw inferences.

In 2018 Pudding Cup, John Miller[9] conducted a textual analysis of country lyrics using topic modelling. He conducted analysis of popularity of words such as "trucks" and "beer" in country music and also performed a time-series analysis of lexical diversity of lyrics in country music.

In this project, I also perform a similar analysis of popular topics and words in pop music. But, I conduct a further analysis of sub-genres and influences of other genres in pop song lyrics. As opposed to a continuous time series analysis, I perform a burst time-series analysis to analyze word repetitions.

Methodological Approach

To solve this problem, we follow given steps:

1. Pre-processing and Cleaning The Data
2. Exploratory Data Analysis of Features of Interest
3. Topic Model Training and Prediction on Test Set
4. Visualization of Results using Word Clouds and Graphs
5. Using aformentioned results for Burst Analysis.

Data Used

The most commonly used datasets for this analysis are musiXmatch dataset[7] that contains song lyrics, stored as bag-of-words representation. Additionally, custom-made datasets are used, which are created by scraping using Spotify API to get artist names and track names and Genius API to get the associated lyrics.

Theory and Hypotheses

In this paper, we have three main objectives:

1. **Predicting the artist's musical style based on their song lyrics** On comparing the most common topics in pop songs to the most popular topic of songs from other genres, we expect to find out the genres that most likely influenced the presence of the topic and its related words in the pop song.

For example, on calculating similarity between 3 most common topics in pop songs and the most common topic of rock music, we find that Topic 1 and 2 are most similar to the rock music topic. We expect the artists whose songs include words from Topic 1 and 2 the most to have a rock/rock-influenced musical style.

2. **Time-series analysis of lyrical trends and sub-genres in pop music** On analysing the increase/decrease in the prevalence of a particular lyric topic and sub-genre in spans of a decade, we expect to find the drivers of increase/decrease such as the evolution of rock music from *rock 'n' roll* to *folk and country-style rock* in the 1960s[3], and the increased popularity of hip-hop as a genre in 2010s.
3. **Burst Analysis of word usage over time** By performing burst analysis, we expect to find two things: increase in intensity of word usage due to events/experience and effect of increased usage of certain words on lyrical creativity.

Data and Methods

About The Dataset

For our analysis, we use the **Music Dataset: Lyrics and Metadata from 1950 to 2019** dataset[1], which contains structural and acoustic features such as valence, danceability, etc. as well as textual features such as lyrics. The dataset contains songs by popular American artists, as well as transliterated lyrics of songs by artists from other countries such as UK, Canada, India, Korea, etc.

We use a subset of the dataset, containing only the columns: *artist_name, track_name, release_date, genre, lyrics, topic.*

To evaluate the performance of the model in predicting an artist's music style, we test the model on a self-created dataset of the top 10 most popular pop songs on *Spotify* in 2020. The dataset contains the above columns, and an additional column **Style**, that represents the true musical style of the artist.

Data Pre-processing

To prepare the data for modelling, we perform the following pre-processing steps:

1. Removal of non-ASCII characters from the “Lyrics” column of the dataset.
2. Creating a new column “Decade” from the “Release Date” Column by rounding off each year to the nearest decade.

Additionally, while creating the Document - Feature Matrix (DFM), we remove punctuations, numbers and stopwords, and convert all words to lowercase.

Exploratory Data Analysis

Before model development, we perform an exploratory analysis on the dataset to get an overview of various feature values and distribution of data across various features.

Analysis of Given Topic Categories

```
        topic
1     sadness
2  world/life
3       music
4    romantic
6    violence
11    obscene
12 night/time
20   feelings
```

The **Topic** column of the dataset has 8 unique values. For appropriate comparison, we also set the number of topics(k) in our model to 8, so that we can compare how similar the extracted topics are to the given topics.

Analysis of Genre

```
# A tibble: 7 x 2
  genre   num_songs
  <chr>       <int>
1 blues        4604
2 country      5445
3 hip hop       904
4 jazz         3845
5 pop          7042
6 reggae       2498
7 rock         4034
```

The **Genre** column of the dataset has 7 unique values. On observing the number of songs in each genre, we observe that maximum songs belong to the *pop* genre, followed by *country* and *blues*.

Pre-liminary Analysis of Pop Song Lyrics

```
# A tibble: 8 x 2
  topic      num_songs
  <chr>          <int>
1 feelings         140
2 music            503
3 night/time       455
4 obscene         1220
5 romantic         431
6 sadness         1702
7 violence        1242
8 world/life      1349
```

On performing a preliminary analysis of the number of pop songs in each topic category, we observe that the most common topic in most songs is **sadness**, followed by **world/life**. We use this as a baseline to compare the results of most common topics predicted by the topic model.

Methods

We use topic modelling to extract the most common topics in pop songs. We also extract the most common topic in other genres using topic modelling and use it to find sub-genres in pop music.

Topic Modelling: LDA To analyse the latent structure of song lyrics, we fit a Latent Dirichlet Allocation(LDA) model with Gibbs sampling for 8 topics(k = 8) and 3000 iterations.

For further analysis of lyrical meaning of pop songs, we observe the top 10 terms for each of the 8 topics.

```
     Topic 1  Topic 2  Topic 3  Topic 4   Topic 5 Topic 6  Topic 7  Topic 8
[1,] "time"   "like"   "like"   "hold"    "head"  "heart"  "know"   "life"
[2,] "feel"   "come"   "yeah"   "fall"    "sleep" "away"   "want"   "live"
[3,] "baby"   "wanna"  "fuck"   "dream"   "blue"  "leave"  "need"   "world"
[4,] "night"  "girl"   "shit"   "hand"    "save"  "go"     "think"  "hear"
[5,] "right"  "better" "bitch"  "eye"     "dead"  "break"  "long"   "home"
```

```
 [6,] "gonna"   "little" "money"  "kiss"     "fight" "walk"    "look"   "sing"
 [7,] "yeah"    "play"   "nigga"  "stand"    "kill"  "stay"    "tell"   "come"
 [8,] "good"    "cause"  "niggas" "remember" "alive" "believe" "change" "song"
 [9,] "mind"    "start"  "cause"  "sweet"    "burn"  "tear"    "things" "place"
[10,] "tonight" "talk"   "bout"   "lonely"   "wall"  "love"    "face"   "bring"
```

Gamma Distribution: Prevalence of Predicted Topics in Given Topics

To create lyrical themes, we look at the per-document per-topic probabilities i.e. the Gamma probabilities. To find the prevalence of each of the 8 given topics on the predicted topics, we estimate the mean contribution for each predicted topic over each given topic.

```
  Given Topics          1         2          3          4          5          6
1     feelings 0.24818181 0.1064151 0.08480770 0.09058292 0.08892938 0.13294081
2        music 0.10712351 0.1169244 0.09099150 0.13374844 0.11609019 0.09872286
3   night/time 0.26034238 0.1108245 0.08740652 0.10465530 0.10137207 0.10213544
4      obscene 0.10728455 0.1608399 0.30182585 0.07203434 0.09124380 0.07439134
5     romantic 0.11672390 0.1168874 0.08065854 0.25380447 0.09050318 0.11859266
6      sadness 0.11625505 0.1122678 0.08071981 0.13971874 0.10685145 0.21421406
7     violence 0.09975353 0.1175207 0.09387288 0.12115696 0.22865737 0.10455538
8   world/life 0.11930440 0.1048538 0.08559031 0.13556664 0.11407772 0.11935515
          7          8
1 0.1254281 0.12271418
2 0.1025488 0.23385031
3 0.1384061 0.09485776
4 0.1045537 0.08782645
5 0.1218726 0.10095724
6 0.1266056 0.10336755
7 0.1160399 0.11844321
8 0.1608994 0.16035249
```

Extraction of Lyrical Themes On observing the average contribution table of given topics and predicted topics, we observe that certain predicted topics are prevalent in certain given topics. For example, Topic 3 is most prevalent in "Obscene" topic category. Additionally, Topic 2 is almost equally prevalent across all given topics. Using these observations, we name the predicted topics the following:

```
  topi_number predicted_topic
1           1     Night/Party
2           2     Base Lyrics
3           3        Explicit
4           4        Romantic
5           5         Violent
6           6     Melancholic
```

```
7           7      About Life
8           8         Musical
```

Finding Sub-Genres in Pop Songs: Influence of Other Genres To find the sub-genres, we first subset the dataset by genre and apply topic modelling to all 7 subsets to extract the most common topic in each genre. We then calculate the **Levenshtein distance** between the 8 topics of pop songs and each genre's most common topic. This gives us the most probable origin genre for a particular topic. For example, the Levenshtein distance between Topic 1 and Hip-Hop Topic is the least. So, we can say that the pop songs and the artists where Topic 1 prevails the most are most likely hip-hop/hip-hop influenced songs.

In this way, we find out the sub-genre for each of the topics.

```
[1] "Night/Party  -  Hip Hop/Reggae" "Base Lyrics  -  Country Pop"
[3] "Explicit  -  Reggae/Hip Hop"    "Romantic  -  Country Rock"
[5] "Violent  -  Country Rock"       "Melancholic  -  Country Pop"
[7] "About Life  -  Rock/Country"    "Musical  -  Rock/Blues"
```

LDA Prediction on Test Set To make predictions on test set, we use the posterior distrubution of the model to predict the most common topics in songs from the test set and extract the maximum value i.e. the most common topic for every data point.

Results

1. Predicting the artist's musical style based on their song lyrics

```
                     artist_name Genre                      Style
1                     The Weeknd   Pop R&B, Electropop, Synth-pop
2                    Roddy Ricch   Pop               Hip-hop/Trap
3  Cardi B, Megan Three Stallion   Pop                    Hip-hop
4            DaBaby, Roddy Ricch   Pop               Hip-hop/Trap
5            24kGoldn, Iann Dior   Pop       Hip-hop,Emo-rap, R&B
6                  Ariana Grande   Pop                   Pop, R&B
7                     Tate McRae   Pop                  Indie Pop
8                      Pop Smoke   Pop    Hip-hop/ Brooklyn drill
9         Bad Bunny, Jhay Cortez   Pop         Latin-trap, reggae
10                      Doja Cat   Pop               Hip-hop, R&B
               Predicted_Style
1  About Life  -  Rock/Country
2  Explicit  -  Reggae/Hip Hop
3  Explicit  -  Reggae/Hip Hop
4  Explicit  -  Reggae/Hip Hop
5  Base Lyrics  -  Country Pop
6  About Life  -  Rock/Country
7  About Life  -  Rock/Country
8  Base Lyrics  -  Country Pop
9  Base Lyrics  -  Country Pop
10 Explicit  -  Reggae/Hip Hop
```

On looking at the musical styles predicted by the model, we observe that the model predicts the style closest to the true musical style of the artist for 7/10 data points.

Looking at the word clouds(Appendix), we observe that the two most common topics in pop songs as predicted by the model are **Violence**, followed by **Melancholy**. This differs from our baseline observation that indicated **Sadness** and **World/Life**.

2. Time-series analysis of lyrical themes and sub-genres in pop music On observing the barplot visualizing the distribution of topics over the years(Appendix), we observe that the occurence of the topic **Explicit** has increased significantly from 2000-2020. The topic of **Romanctic** dominated the 1960s but its occurence has significantly reduced in 2010-2020.

In the 1960s, the proportion of **Romantic - Country Rock** is higher. This reflects the evolution of rock music from *rock 'n' roll* to a more folk/country-style rock music. Since **Romantic** is a the most common topic in country music, the pop songs in 1960s were a fusion of country and rock music.

The sharp rise in the proportion of **Explicit - Hip-Hop/Reggae** in 2010s and 2020s reflects the breakthrough of hip-hop and rap music in the **Billboard Hot 100** chart. It marks the rise of artists like Kendrick Lamar, Drake, Nicki Minaj, etc to fame and also the increased fame of rap producers and DJs.

Hip-hop/Reggae as a sub-genre of mainstream pop music shows a continual increase from 1950 t0 2020. Looking at the history of hip-hop[4], it originated in 1970s New York's The Bronx as a cultural movement among African-Americans, Puerto-Ricans and Carribean immigrants. It reflected the harsh economic and political conditions of the city and acted as an outlet for grief, anger and expression of hardships and struggles faced by people due to being immigrants.

As years went by, more and more generations of immigrants started using hip-hop and rap music as a medium to tell their stories with OG rappers like DJ Kool Herc that inspired a generation of Jamaican young adults to express their sentiments through rap and music. Hence, we can see the rise in the proportion of **Explicit - Hip-Hop/Reggae** topic in pop songs over the years.

3. Burst Analysis of word usage over time

Looking at the level plot for *heart*(Appendix), we observe that the intensity of usage of the word **heart** increased from 1956 to 1963. Following that period, the intensity level dropped until 2015 where it increased again for a year. To analyze the increase in 1956-1963, we look at the artists with the most releases in that period.

```
# A tibble: 5 x 2
  artist_name   num_docs
  <chr>            <int>
1 sam cooke           36
2 johnny mathis       33
3 roy orbison         30
4 the platters        29
5 neil sedaka         25
```

On looking at the musical styles of the artists, we observe that their styles include R&B, Soul, Brill Building and Country. These releases reflect the 1950s and 1960s where genres like R&B and jazz became more mainstream and rise to prominence of British blues and folk music.

On observing the level plot for the word **yeah**(Appendix), we observe that its usage increased to a level 2 in the 2000s and increased to level 3 in 2018-2019. This bolsters the claim of studies that say that pop song lyrics have become repetitive in the recent years. With the increased need to make songs "catchy", composers tend to add "yeah yeah yeah" and similar terms to songs. This trend is also popular in K-pop songs where the English translation follows the Korean lyrics.

Since this trend has emerged in the recent years, one can say that to make a song catchy and popular, songwriters compromise with creativity and use words like "yeah" as fillers.

Discussion

By performing a lyrical analysis of pop songs, we aimed to gain insights into the human mind and the society we live in and the role pop music plays in developing/degrading creativity and development of personality and belief system of young adults.

We used topic modelling to predict an artist's musical style based on their lyrics and conduct a time-series analysis of trends in pop music related to sub-genres and influence of other music genres. We observed that the model predicts on unknown data with 70% accuracy. To improve the performance of the model, the dataset needs to be expanded to include songs from other, lesser known genres such as soul, R&B, folk, etc. Additionally, transliterating from other languages may alter the original meaning of the lyrics, depending upon the accuracy of transliteration. Hence, a multi-lingual textual analysis needs to be performed to get accurate predictions.

My contributions include: using LDA to create sub-genres in pop music based on song lyrics which can be used to improve pop song recommendations based on a user's musical taste. It can also be used to curb recommendations of explicit or inappropriate songs to minors. Secondly, I performed a burst analysis of usage of certain words. The results reflect the change in musical trends over the years, the rise and fall in popularity of certain genres and also creativity in lyrical content. I also conducted a time-series analysis that reftected the increase/decrease in proportion of a particular topic in each decade.

References

[1] Moura, Luan; Fontelles, Emanuel; Sampaio, Vinicius; França, Mardônio (2020), “Music Dataset: Lyrics and Metadata from 1950 to 2019”, Mendeley Data, V2, doi: 10.17632/3t9vbwxgr5.2

[2]Pereira, Carlos Silva, et al. “Music and emotions in the brain: familiarity matters.” PloS one 6.11 (2011): e27241.

[3] https://en.wikipedia.org/wiki/1960s_in_music

[4] https://iconcollective.edu/hip-hop-history/

[5] Cha, Kyoung Cheon, et al. “Young consumers’ brain responses to pop music on Youtube.” Asia Pacific Journal of Marketing and Logistics (2019).

[6] Kreyer, Rolf, and Joybrato Mukherjee. “The style of pop song lyrics: A corpus-linguistic pilot study.” (2007): 31-58.

[7] Thierry Bertin-Mahieux and Daniel P.W. Ellis and Brian Whitman and Paul Lamere. Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011) [8] https://github.com/johnwmillr/trucks-and-beer

Appendix

Figure 1: Word Clouds of Top 2 Topic in Pop Songs

```
  Topic Number Frequency
5            5      1100
6            6      1093
```

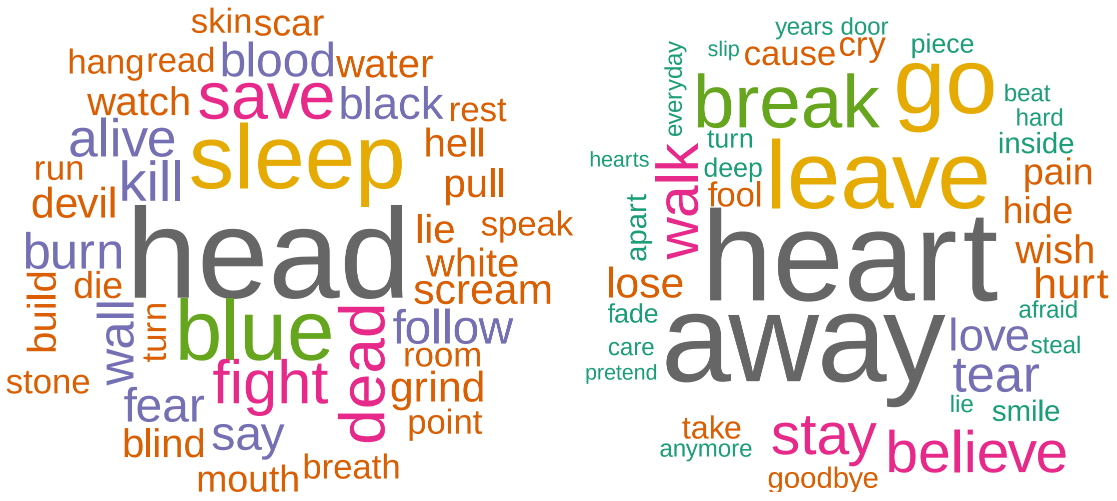


Figure 2: Time-series analysis of lyrical themes and sub-genres in pop music

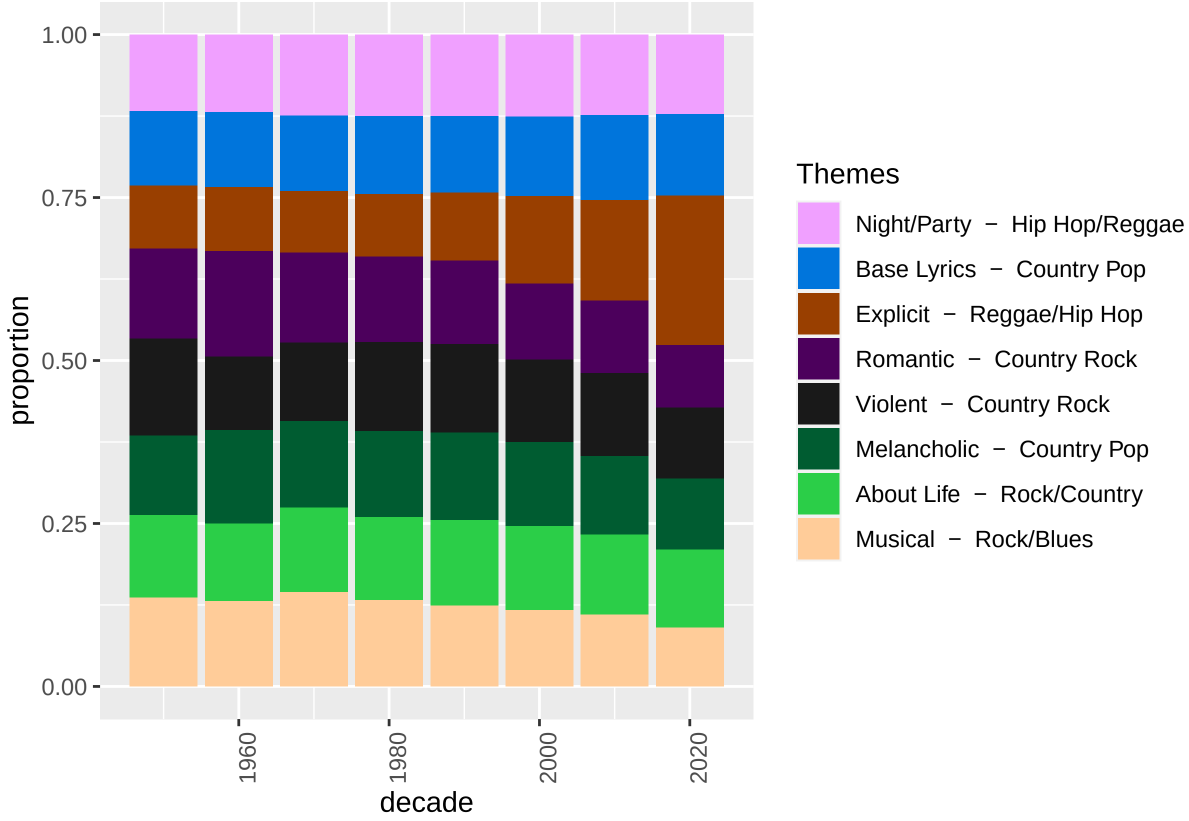
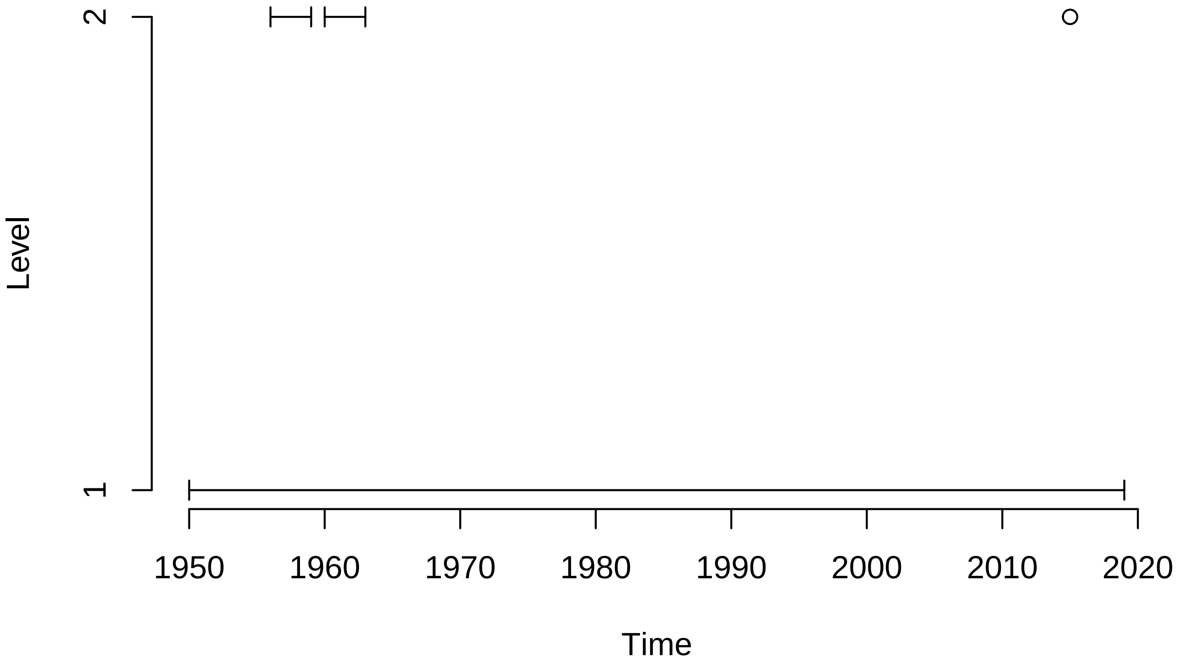
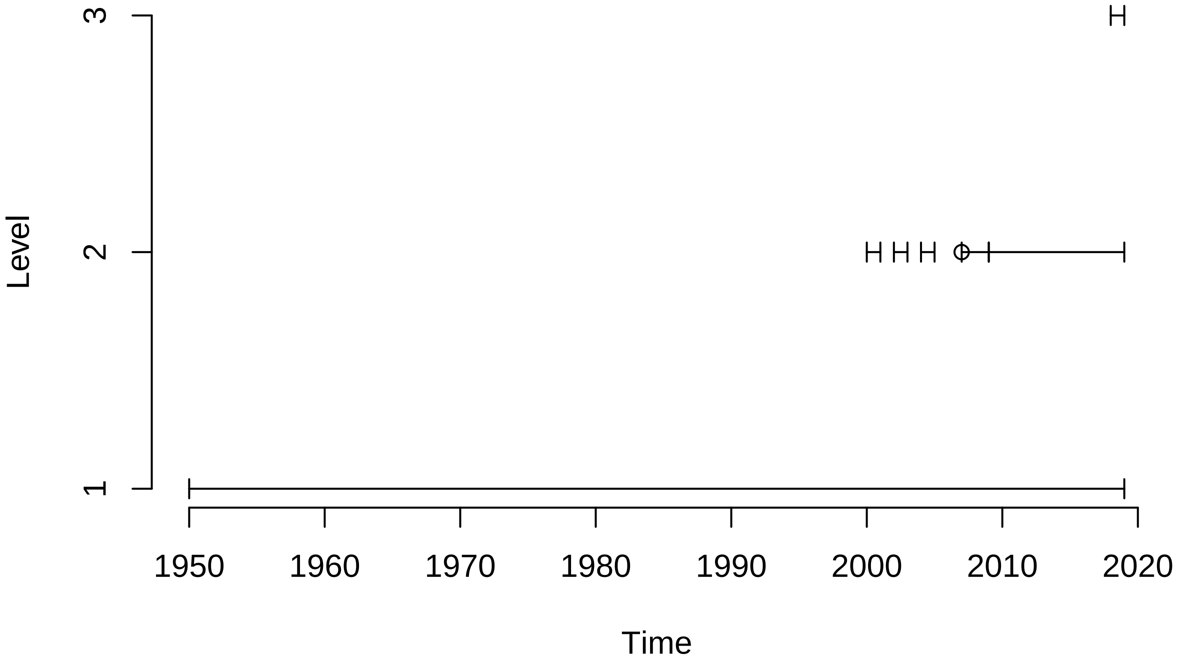


Figure 3: Burst Analysis of “heart”



```
  level start  end
1     1  1950 2019
2     2  1956 1959
3     2  1960 1963
4     2  2015 2015
```

Figure 4: Burst Analysis of “yeah”



```
  level start  end
1     1  1950 2019
2     2  2000 2001
3     2  2002 2003
4     2  2004 2005
5     2  2007 2007
6     2  2007 2009
7     2  2009 2019
8     3  2018 2019
```