TikTok Trending Songs

Authors: Kenny Duong, Erick Gomez, Karina Gonzalez, Jammal Loiseau, Javier Machuca
Department of Information Systems, California State University Los Angeles
CIS4560-01 System Analysis and Design
kduong31@calsatela.edu, egomez103@calstatela.edu, kgonz163@calstatela.edu, jmachu13@calstatela.edu

Abstract

The paper explains the method and process used for extracting the data from scraped videos from TikTok and figuring out what songs are most popular. In addition to that, analysis of this data is conducted using Excel and shows us what songs are used most often in terms of the success of a video on the platform.

1. Introduction

This project uses Hadoop and Hive to keep and process TikTok trending videos dataset. The dataset is mainly music artists, the song name, album name, disc number, title name, release date, record label, video length, source of the music. We have chosen this dataset because TikTok is one the most popular social media platforms, and we wish to find exactly what trends contribute to a viral video. There are many factors at play, but we believe that music plays a vital role in determining the outcome/success of a video. Based on the data we have gathered; we strongly believe that videos have a better chance of going viral if they use a popular soundtrack.

2. Related Work

Although TikTok is a popular social media platform to share videos, there are some publications that show how it is also becoming a popular site to share and discover new music. The work from MBW (Music Business Insider World) reveals that there is an increasingly higher number of monthly active users (MAUS). There are estimates that indicate it will only increase as time goes on.

There was one work based around music data. The publication referenced by MRC Data. They are the provider of music related data, and they explain that TikTok is on its way to become a strong catalyst for artists to present them music to a wider audience. Data from the study shows that a majority of TikTok users use the app to share new artists and aid in discoverability. Another insight the report found is that TikTok users are more likely to seek out a song that they heard while watching a video. The study also suggests that the original content that is posted by users, followed by a song, will encourage others to further seek out the artist. The data they have obtained supports our theory on how music contributes to the growth of a TikTok video. The data explains the impact music has, and in turn how viral it is able to be. For our data, we will take a closer look at the data that consists of individual songs used in all the video data. We will be able to take an in-depth look at how certain artists will fare against each other.

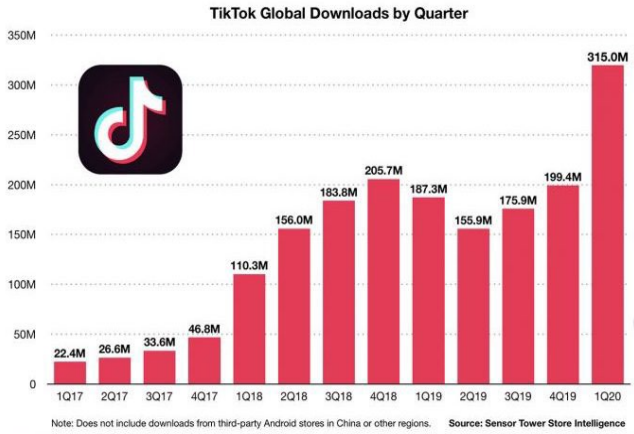

Figure 1 - TikTok Global Downloads (quarterly)

3. Background

When it comes to the app TikTok it has become one of the most premiere social media platforms in the world. TikTok is the successor to another short video social platform, Vine. After the shutting down of Vine, the creators decided to go back to the drawing board and recreate and rebrand the platform with TikTok. The reason we choose this topic is because as alluded to before, seventy five percent of the users on TikTok say that they have discovered new artists according to Murray Stassen of MBW. The article goes on to elaborate on how TikTok has been able to "reveal the power of music" through its platform allowing it to influence and shed light on all types of backgrounds and communities. This development caused our group to wonder about the most popular and trending videos on the platform and what songs were featured in those videos.

4. Specifications

The dataset comprises of the id number, artist, album, release date, label, and timestamp, song link, the ISRC [1] of Apple Music, and the ID of the song on Spotify. The dataset has a size of 192 KB (not counting the size of the videos themselves) and 3 GB in total. The dataset spans several files with each containing metadata from different music sources such as Spotify, Apple Music. This dataset contains data scrapped of trending videos at that time in 2020. The trending videos results are curated based on the user's personal account so the results may vary slightly. Table 1 shows files and the size of the files

Table 1 Data Specification

Data Set	Size (Total 752 KB)
audd_music	84 KB
audd_music_apple_music	89 KB
audd_music_spotify_music	186 KB
audd_music_spotify_music_artists	393 KB

The below table shows the specification for Oracle cluster we are using and the Hadoop specification for our project.

Table 2 H/W Specification

Number of nodes	1
OCPUs	3
CPU speed	2400 MHz
Memory	35187484 B
Storage	144 GB

1 **ISRC**, the **International Standard Recording Code**, is the internationally recognized identification tool for sound and music video recordings.

The dataset is composed of the various songs used and their respective ID numbers. The whole process we used from downloading the dataset to how we manipulated the data is shown below in the graphic. (Figure 1). There are four files in total each focusing on different sources of music (Spotify and Apple Music etc.) After that, HiveQL was used as the querying language to create the tables' schema. We then created a summary table to find the top trending videos and export the results. The output file was downloaded and viewed as an Excel file. We then analyzed the data to create visualizations.

5. Implementation Flowchart

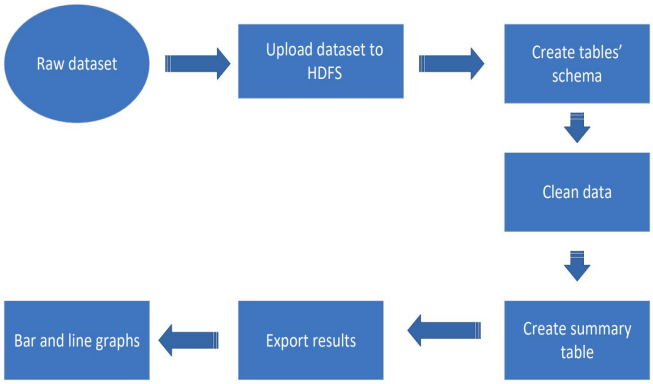

Figure 2 - Implementation Flowchart

The flowchart starts with the raw dataset we obtained from Kaggle. Next, we uploaded the dataset to HDFS. Next, we created a table schema, but the data required some cleaning. This part required more time as there were some null values that had to be excluded from the result. Once the data was properly cleaned, we could create a summary table.

We would base our results on this newly created table. We exported the results. With Tableau, we were able to create visualizations representing factors such as popularity and time based factors like duration.

6. Data Cleaning

Audio files were uploaded and stored in HDFS and then loaded into tables using Hadoop's Beeline Client. The dataset that we acquired required minimal cleaning. When we analyzed the files, the columns were already categorized which made it straightforward to work with.

Using Git Bash to clean and engineer the data to our liking we found a tremendous number of errors. We tried altering the tables so that they could be joined, but that caused values to be nulled even though the data types remained the same. We decided to use the AS SELECT function to grab attributes from the tables that we wanted. Then we used INSERT to download the cleaned tables back to the local Linux machine into new directories for each table. From then on, we moved on to Tableau for visualization.

7. Analysis and Visualization

Once the data was cleaned and prepared, we completed our analysis by loading the data into hive.

After data cleaning and preparation for further analysis, files were extracted into Excel and through Tableau. We used Tableau to create two graphs, one displaying Popularity and another based on Duration of each song.

The first visualization (Figure 2), a bar graph, was made in Tableau and each color represents a different song. Top 10 songs are shown below. The top song is 'Mood' by 24KGoldn ft. iann Dior followed by 'Without You' leading in second.

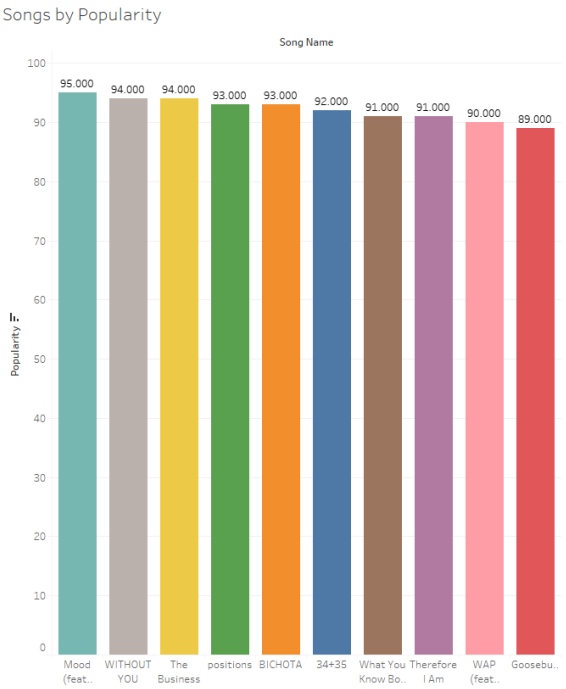

Figure 3 - Song by popularity sorted by top 10.

7.1 Time Based visualization (duration)

The third visualization represents the duration of each song. The top song is called 'TikTok' which was particularly interesting. We have dived deeper into this and found that that the most popular song at the time was classified as this name. The song could be any trending song at that moment in time. One of the songs most likely included 24KGoldn ft iann Dior which is the top result in our popularity graph. The total duration of this song is about 8 min followed by 'Without You' leading in second. The duration is measured in milliseconds.

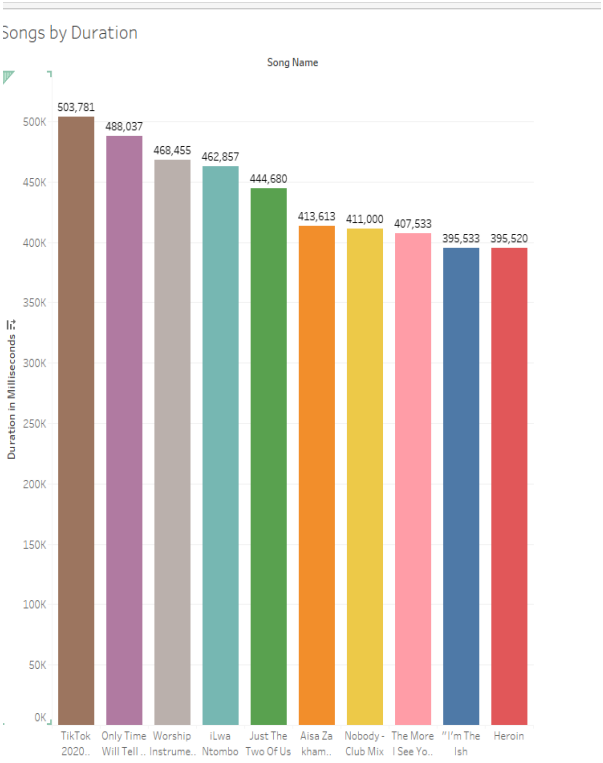

Figure 4 – Songs by Duration sorted by the top 10

7. Conclusion

Finally, summing up all the above work we can conclude the following:

I. There is a correlation between a song's popularity and its ability to help a video go viral.

II. From the data, we can conclude that videos will have more success if they are associated with a catchy tune.

III. We can also verify that the songs are popular because they are among the top ten most-heard songs, demonstrating that popular music aids in the spread of video trends.

To create the information above, we used interactive technologies like HDFS and Beeline to manipulate the data. We then used Tableau to create visualizations to better analyze the cleaned data we created. For more information and the code, visit the project's GitHub link 3.

8. References

TikTok Dataset (2021) by Erick Van de Ven. Retrieved from https://www.kaggle.com/datasets/erikvdven/tiktok-trending-december-2020

75% of TikTok users say they Discover New Artists on the Platform - Music Business Worldwide. Retrieved from https://www.musicbusinessworldwide.com/tiktok-has-over-800m-active-users-worldwide-75-of-them-say-they-discover-new-artists-on-the-platform/

What is ISRC? ISRC. (n.d.). Retrieved May 21, 2022,

[3] GitHub Link: https://github.com/JamBaby23/TikTok/