Datasheet for the Dataset

- Ameya Naik (Author of this datasheet and curator of the dataset)

Motivation for Dataset Creation

- Why was the dataset created? The Artist Song Lyric Dataset was specifically created by Ameya Naik for the purpose of implementing an analysis of closeness of artists on the basis of their lyrics. So it contains top 100 songs as per their popularity at genius.com of selected artists.
- What (other) tasks could the dataset be used for? The dataset could be used to analyze the lyrical usage by a particular artist. This data set could also provide good information about who similar top 100 popular songs are.Moreover the dataset folder also contains metadata about the artist which could be used to derive further analysis of artist genre , style with the types of lyrics.
- Has the dataset been used for any tasks already? This github repository previews the original intention of creating the dataset and the usage.
- Who funded the creation of the dataset? The dataset was just created by the aforementioned author using the Genius API for academic purposes and could be further used for personal projects. Hence there was no external funding provided or needed.

Dataset Composition

- What are the instances? The dataset contains the text file consisting of the top 100 song lyrics, each text file is named after the artist. Along with the artist names on the text files as well as an artist metadata file.
- Are relationships between instances made explicit in the data? There's no relationship as such between the artists except the artist found those artists to be interesting to work up.
- How many instances of each type are there? Each instance is a song lyrics file by the artists so the number of instances we have would vary depending on the artist we include. Presently during documenting this datasheet there are 25 artists I have added in the datasets.
- What data does each instance consist of? All the top 100 songs lyrics in a text file.
- Is everything included or does the data rely on external resources? The data set contains everything included.
- Are there recommended data splits or evaluation measures? The dataset if for analysis and visualization hence there isn't specifically train-test split or other evaluation measures. One could check the number of distinct word being used in all the 100 songs by the artist for reference.

- What experiments were initially run on this dataset? The dataset was added without any experimental results however the this github(https://github.com/ameya-9/Lyrics-Analysis) project provides some analysis that could be run.
- How was the data associated with each instance acquired? The name of each music artist was the name of the text file having their songs lyrics, it's in text format.
- Does the dataset contain all possible instances? No the dataset is not exhaustive, rather just list lyrics for the artist which the author was interested in. This could be further expanded by adding more artists.
- If the dataset is a sample, then what is the population? The population for this dataset would be all the artists who have their songs lyrics on the Genius.com websites. So since currently it's only 25 artists we are looking into, this would be expanded to 100s depending upon the researcher requirements.
- Is there information missing from the dataset and why? Yes there were few instances where the songs lyrics couldn't be retrieved from the website due to unavailability. In such a situation the next ranked songs were retrieved so as to have top 100 songs which could be fetched.
- Are there any known errors, sources of noise, or redundancies in the data? Nope

Data Preprocessing

- What preprocessing/cleaning was done?

  The following steps were taken to process the data (both the lyrics and artist data).

  1. Getting the list of artists we are interested in: After discussion with a few of my colleagues and friends, I devised a list of artists that would be interesting.
  2. Gathering the lyrics: I have used Genius API (created my developer profile and did the necessary steps for accessing the data). Then using the genius api, I retrieved the lyrics of the top 100 songs of the previously created list of artists.
  3. Gathering data about artists : To gather data about artists, I referred to the kaggle dataset, sliced and diced the data as per my requirements and wrote it on a new csv file.
  4. During Doc2vec creation: For the lyrics dataset since I wanted to gather interesting, I had remove the filler words, curse words as well as stop words, and all the other non-english words(that are part of flow, or just to rhyme).Also converting everything to lowercase for standard usage.
- Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet ? Yes, the only limitation is currently the dataset contains only top 100 songs rather all the possible available songs, which understandably is difficult to retrieve.

Dataset Distribution

- How is the dataset distributed? The data set can be downloaded from my github repository, the details of retrieval of the datasets are in this document. Moreover the code to use Genius API is available in the src folder of my github profile.(with proper links to original sources).The images can be downloaded as a zipped bundle of text files.

- When will the dataset be released/first distributed? The dataset was released on 15th December, 2021.
- What license (if any) is it distributed under? Are there any copyrights on the data? The crawled data from Genius belongs to the website, however could be used for academic and personal project purposes.
- Are there any fees or access/export restrictions? There are no fees or restrictions.

Dataset Maintenance

- Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset? The dataset is hosted on github repository of Ameya Naik. Incase of any comments or emails please contact me via my email: ameyanaik9@gmail.com.
- Will the dataset be updated? Yes the dataset would be updated by the author, as the number of songs needs to be updated as well as to include newer artists.
- If the dataset becomes obsolete how will this be communicated? The details would be posted on the readme file of the repository for users to be aware of.
- If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? The author intends to share the dataset on Kaggle after confirmation of the legal restrictions, in such situations it would be easier to update the datasets.
  For now, the user could do their required changes as per their requirement in the code to retrieve more data, this could be done by forking the code “LyricsFromGeniusAPI.ipynb” in src folder.

Legal & Ethical Considerations

- If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? Since the dataset was retrieved from the website using the authentication code and token, the original curator is aware about the information that was retrieved from the database and by which user.
  A
  As the dataset is about artists and their song lyrics which are publicly available hence there’s issue of confidentiality or privacy breach.
- If it relates to other ethically protected subjects, have appropriate obligations have been met? (e.g., medical data might include information collected from animals)Not applicable
- If it relates to people, were there any ethical review applications/ reviews/approvals?Not applicable
- If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications?No (refer first question).

- If it relates to people, could this dataset expose people to harm or legal action? The artist information were publically available as this doesn't pose minimal to no risk to them.[1]
- Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information) The dataset does not contain confidential information since all information was scraped from websites.
- Does the dataset contain information that might be considered inappropriate or offensive? Yes, some of the lyrics and words could be considered inappropriate for individuals below a certain age, this has been taken into consideration while doing analysis.

[1] Datasheets for Datasets: https://www.fatml.org/media/documents/datasheets_for_datasets.pdf
Timnit Gebru 1 Jamie Morgenstern 2 Briana Vecchione 3 Jennifer Wortman Vaughan 1 Hanna Wallach 1
Hal Daumé III 1 4 Kate Crawford 1 5