Dataset Description

The lyrics-data.csv file from the Kaggle dataset "Song lyrics from 79 musical genres" is a comprehensive collection of song lyrics data scraped from the Brazilian website Vagalume.com.br.

Website Access

**URL (Kaggle Dataset Link)**: <https://www.kaggle.com/datasets/neisse/scrapped-lyrics-from-6-genres?select=lyrics-data.csv>

Dataset Overview

* **Name:** lyrics-data.csv
* **Source:** Vagalume.com.br
* **Scraped by:** Anderson Neisse
* **Last Updated:** 2 years ago
* **Size:** 434.3 MB
* **Number of Songs:** 379,893
* **Number of Artists:** 4,239
* **Languages:** Primarily English (50%) and Portuguese (41%), among others.

Language distribution

Here's a summary of the dataset's language distribution:

* **English (en):** 191,814 songs
* **Portuguese (pt):** 157,393 songs
* **Spanish (es):** 9,917 songs
* **Italian (it):** 1,432 songs
* **French (fr):** 1,225 songs
* **German (de):** 844 songs
* **Kinyarwanda (rw):** 1,679 songs
* **Finnish (fi):** 145 songs
* **Swedish (sv):** 112 songs
* **Romanian (ro):** 97 songs
* **Norwegian (no):** 89 songs
* **Icelandic (is):** 86 songs
* **Afrikaans (af):** 19 songs
* **Sundanese (su):** 19 songs
* **Swahili (sw):** 19 songs
* **Welsh (cy):** 23 songs
* **Indonesian (id):** 26 songs
* **Galician (gl):** 36 songs
* **Irish (ga):** 32 songs
* **Turkish (tr):** 32 songs
* **Catalan (ca):** 13 songs
* **Danish (da):** 13 songs
* **Estonian (et):** 13 songs
* **Dutch (nl):** 14 songs
* **Korean (ko):** 17 songs
* **Tagalog (tl):** 69 songs
* **Polish (pl):** 47 songs
* **Arabic (ar):** 4 songs
* **Basque (eu):** 4 songs
* **Scottish Gaelic (gd):** 4 songs
* **Russian (ru):** 4 songs
* **Haitian Creole (ht):** 5 songs
* **Japanese (ja):** 7 songs
* **Sesotho (st):** 7 songs
* **Malay (ms):** 8 songs
* **Czech (cs):** 3 songs
* **Kurdish (ku):** 3 songs
* **Malagasy (mg):** 3 songs
* **Chichewa (ny):** 3 songs
* **Hungarian (hu):** 2 songs
* **Javanese (jw):** 2 songs
* **Luganda (lg):** 2 songs
* **Albanian (sq):** 1 song
* **Croatian (hr):** 1 song
* **Hebrew (iw):** 1 song
* **Latvian (lv):** 1 song
* **Serbian (sr):** 1 song
* **Slovenian (sl):** 1 song
* **Vietnamese (vi):** 1 song
* **Chinese (zh):** 1 song
* **Persian (fa):** 1 song

This distribution indicates a strong presence of English and Portuguese songs, which is reflective of the source website's primary audience. The dataset also includes a variety of other languages, though with significantly fewer songs, which could be indicative of the global reach of the website or the popularity of certain genres within those language groups. For data science purposes, this diversity allows for comparative studies across languages and the development of language-specific models.

Data Attributes

* **ALink**: The URL to the artist's profile on Vagalume.com.
* **SName**: The name of the song.
* **SLink**: The URL to the song's lyrics on Vagalume.com.
* **Lyric**: The full text of the song lyrics.
* **Language**: The language code for the lyrics.

Data Structure Analysis

**Format**: CSV (Comma-Separated Values)

**Nature of Data**: Semi-structured. While the CSV format is a structured format, the lyrical content within the 'Lyric' field is inherently unstructured text data.

**Complexity Level**: Given the size of the dataset and the unstructured nature of text data, the complexity level is high. The dataset requires advanced text processing and natural language processing techniques to structure and analyze the lyrical content.

Data Content Analysis

**Genre Diversity**: Encompasses a broad spectrum of 79 musical genres.

**Artist Representation**: Includes a wide array of 4,239 artists, indicating a diverse sampling of musical styles and expressions.

**Lyrical Depth**: Provides a rich corpus for deep linguistic and thematic analysis.

Potential Data Science Use Cases

**Lyric Analysis**: To study linguistic features, sentiment, and themes.

**Natural Language Processing (NLP)**: For training models on tasks such as text generation, language detection, and machine translation.

**Cultural Studies**: To gain insights into cultural expressions in music across different regions and languages.

Data Quality Considerations

**Consistency**: Potential inconsistencies due to the nature of web scraping.

**Completeness**: The dataset may have missing information due to the limitations of the scraping process.

**Accuracy**: The accuracy of the lyrics and associated metadata may vary and should be verified for critical applications.

Licensing, Usage, and Ethical Considerations

**License**: Open Database License for the database and Database Contents License for the contents.

**Usage Rights**: Users of the dataset provided by Anderson Neisse / Kaggle are free to share, modify, and use the dataset for various purposes, adhering to the terms of the Open Database License.

**Ethical Considerations**: The use of the dataset should consider the copyright of the lyrical content as well as the privacy of the artists.

Accessibility and Maintenance

**Download Requirements**: Requires a Kaggle account for access.

**Update Frequency**: No regular updates, last updated 2 years ago.

**Support and Maintenance**: Limited, as it is a scraped dataset with no active maintenance.

Sample data entries for illustration

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*To add\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

References