Data Cleaning: Genius Song Lyrics Subset

Purpose: Clean and preprocess song lyrics for analysis. Remove metadata tags (e.g., `[Intro]`, `[Verse]`), line breaks (`\n`) and extra whitespace to create a clean text column suitable for NLP and statistical analysis.

---

1. Dataset Overview

1.1 Import Packages and Settings

In [1]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import re
import os
```

In [2]:

```python
plt.style.use('default')
plt.rcParams['figure.figsize'] = (12, 6)
%matplotlib inline
```

1.2 Load Dataset

In [8]:

```python
df = pd.read_csv('data/raw/lyrics_subset_1pct.csv')
print(f"DataFrame shape: {df.shape}")
print(f"Number of Songs: {len(df)} | Artists: {df['artist'].nunique()}")
df.head()
```

```
DataFrame shape: (51348, 11)
Number of Songs: 51348 | Artists: 39461
```

Out[8]:

	title	tag	artist	year	views	features	lyrics	id	language_cld3	language_ft	language
0	2 Is Better	rap	Chris Travis	2017	4437	{}	[Intro]\nBitch I'm clean\nTwo sticks like Chow...	3036329	en	en	en
1	Scottie	rap	KrJ	2012	89	{}	My old girl left me on her old bull shit\nSo I...	72180	en	en	en
2	Pirate Password	rock	The never land pirate band	2011	175	{}	[Intro: spoken]\nAvast there matey haha\nIf a ...	2122100	en	en	en
3	Indri	rock	Puta Volcano	2015	14	{}	Just throw a glimpse under the shell\nGhostly ...	6889288	en	en	en
4	Maps	misc	ANBARDA	2018	4	{}	[Verse 1]\nI miss the taste of a sweeter life\...	3735887	en	en	en

2. Data Cleaning

2.1 Problem

A preview of the raw lyrics highlights issues such as metadata tags (e.g., `[Intro]`, `[Verse]`) and line breaks (`\n`) that need to be cleaned before analysis.

In [9]:

```
df['lyrics'].head(5)
```

Out[9]:

```
0    [Intro]\nBitch I'm clean\nTwo sticks like Chow...
1    My old girl left me on her old bull shit\nSo I...
2    [Intro: spoken]\nAvast there matey haha\nIf a ...
3    Just throw a glimpse under the shell\nGhostly ...
4    [Verse 1]\nI miss the taste of a sweeter life\...
Name: lyrics, dtype: object
```

2.2 Define and Apply Cleaning Function

- `re.sub(r'\[.*?\]', '', text)` removes everything between [and]
- `text.replace('\n', ' ')` replaces line breaks with a space
- `re.sub(r'\s+', ' ', text).strip()` ensures there are no multiple spaces left and trims the text cleanly

In [10]:

```python
def clean_lyrics(text):
    text = re.sub(r'\[.*?\]', '', text)
    text = text.replace('\n', ' ')
    text = re.sub(r'\s+', ' ', text).strip()
    return text
```

In [11]:

```python
df['lyrics_clean'] = df['lyrics'].apply(clean_lyrics)
```

2.4 Preview Cleaned Lyrics

In [12]:

```python
df[['lyrics','lyrics_clean']].head(10)
```

Out [12]:

	lyrics	lyrics_clean
0	[Intro]\nBitch I'm clean\nTwo sticks like Chow...	Bitch I'm clean Two sticks like Chow Mein Two ...
1	My old girl left me on her old bull shit\nSo I...	My old girl left me on her old bull shit So I ...
2	[Intro: spoken]\nAvast there matey haha\nIf a ...	Avast there matey haha If a pirate asks ya for...
3	Just throw a glimpse under the shell\nGhostly ...	Just throw a glimpse under the shell Ghostly v...
4	[Verse 1]\nI miss the taste of a sweeter life\...	I miss the taste of a sweeter life I miss the ...
5	Shadows been word\n\nOne win a hand\n\nTo sell...	Shadows been word One win a hand To sell it th...
6	Cold sweat dreaming, am I awake? I feel draine...	Cold sweat dreaming, am I awake? I feel draine...
7	[Verse 1]\nOut here walking 'round this empty ...	Out here walking 'round this empty town Got a ...
8	[Verse 1]\nImmer näher an die Sonne, kaum Trei...	Immer näher an die Sonne, kaum Treibstoff mehr...
9	[Hook]\nI waited for too long, I waited, I wai...	I waited for too long, I waited, I waited I wa...

Metadata tags (e.g., `[Intro]`, `[Verse]`), line breaks (`\n`) and extra whitespace are now removed.

3. Save cleaned Data

3.1 Configuration

- Define output directory and file name.
- Create directory if it doesn't exist.

In [13]:

```
output_dir = "data/clean"
os.makedirs(output_dir, exist_ok=True)

output_path = os.path.join(output_dir, "lyrics_subset_1pct_clean.csv")
```

3.2 Prepare Data for Saving

Drop the original `lyrics` column and rename `lyrics_clean` to `lyrics`.

In [14]:

```
df = df.drop(columns=['lyrics'])
df = df.rename(columns={'lyrics_clean': 'lyrics'})
```

3.3 Save Subset locally

In [15]:

```
df.to_csv(output_path, index=False)

print(f"Cleaned Subset saved to: {output_path}")
```

```
Cleaned Subset saved to: data/clean\lyrics_subset_1pct_clean.csv
```

In [ ]:

```
```