

Zusammenfassung des aktuellen Stands des Lyrics-Scraping-Skripts

Kernfunktionalitäten:

1. Metadaten-Extraktion:

- Liest ID3-Tags aus Audiodateien (MP3, FLAC, WAV, M4A, OGG)
- Extrahiert Künstler, Titel und Album-Informationen
- Fallback-Mechanismus bei fehlenden Tags

2. Intelligente URL-Generierung für Genius:

- Zwei URL-Varianten für bessere Trefferquote:
 - Standard: {künstler}-{titel}-lyrics
 - Feature-Version: {künstler and featured}-{titel}-lyrics
- Automatische Umlaut-Konvertierung (ä → ae)
- Sonderzeichen-Entfernung

3. Robustes Scraping mit Fallback-Strategie:

1. Direkter Zugriff via generierter URL
2. Genius-Suche bei Fehlschlag
3. Songtexte.com als letztes Fallback

4. Deutschrap-spezifische Optimierungen:

- "Feat."-Handling: Ersetzt durch "and" für URLs
- Entfernt Tracknummern (z.B. "06.")
- Bereinigt Klammern und Zusätze

5. Dateiverwaltung:

- Speichert Lyrics im selben Ordner wie die Audiodatei
- Dateinamenformat: {Originaldateiname}_lyrics.txt
- Enthält Metadaten-Header in der TXT-Datei

Technische Umsetzung:

• Benötigte Bibliotheken:

```
requests, beautifulsoup4, mutagen, tqdm, unicodedata
```

• Kommandozeilenaufruf:

```
python lyrics_scraper.py --input "Pfad/zum/Musikordner"
```

• Verarbeitungsablauf:

1. Rekursives Durchsuchen des Eingabeordners
2. Verarbeitung jeder Audiodatei mit Fortschrittsbalken
3. Automatische Verzögerung zwischen Anfragen (1.5s)

Beispieldateistruktur:

```
Musikordner/  
├── Artist - Song.mp3
```

```
└─ Artist - Song_lyrics.txt  # Automatisch generiert
   Subordner/
     └─ Another Song.flac
        └─ Another Song_lyrics.txt
```

Wichtigste Funktionen:

1. `clean_rap_metadata()`: Bereinigt Titel/Künstler für Deutschrap
2. `normalize_feature_artists()`: "Feat." → "and" Konvertierung
3. `scrape_genius_lyrics()`: Haupt-Scraping mit zwei URL-Varianten
4. `genius_search_fallback()`: Suchfallback mit Album-Kontext
5. `process_directory()`: Hauptverarbeitung mit tqdm-Fortschrittsbalken

Nächste Schritte für KI-Tagging:

1. Integration der Ollama-API:

- Anbindung an lokale Ollama-Instanz (<http://localhost:11434>)
- Verwendung des DeepSeek-R1-Modells

2. Tag-Generierung:

- Automatische Erstellung von 10 Musik-Tags pro Song
- Berücksichtigung von Genre, Stimmung, Instrumentierung
- Deutschrap-spezifische Tags (Flow-Typ, Beat-Subgenre)

3. Dateiausgabe:

- Neue Datei: {Originaldateiname}_tags.txt
- Optional: Kombination mit Lyrics in einer Datei

4. Workflow-Integration:

- Tagging nach erfolgreichem Lyrics-Scraping
- Fehlerbehandlung bei nicht verfügbarer KI