1. Datenanalyse (Exploratory Data Analysis - EDA)

**Verteilung der Sprachen:**

Schaut euch an, welche Sprachen im Datensatz am häufigsten vorkommen.

Gibt es Unterschiede in der Länge der Liedtexte zwischen den Sprachen?

**Beliebtheit der Künstler und Lieder:**

Welche Künstler haben die meisten Aufrufe (views)?

Gibt es einen Zusammenhang zwischen dem Erscheinungsjahr und der Beliebtheit?

Gibt es einen zusammenhang zwischen der Sprache und der Beliebtheit?

**Analyse der Liedtexte:**

Welche Wörter kommen am häufigsten vor?

Gibt es typische Wortmuster für bestimmte Genres (tag)?

Wie lang sind die durchschnittlichen Liedtexte?

2. Machine Learning

**Genre-Vorhersage (tag):**

Trainiert ein Modell, das das Genre eines Liedes anhand des Textes vorhersagt.

Verwendet dafür zum Beispiel Algorithmen wie Naive Bayes, Support Vector Machines oder Random Forests.

**Beliebtheitsvorhersage (views):**

Versucht, die Anzahl der Aufrufe eines Liedes vorherzusagen, basierend auf dem Text, dem Künstler und dem Erscheinungsjahr.

Hierfür könnt ihr Regressionsmodelle verwenden.

**Spracherkennung (language):**

Obwohl die Sprachen im Datensatz schon vorhanden sind, könnt ihr versuchen, ein eigenes Modell zur Spracherkennung zu trainieren.

Dies könnte ein interessanter Vergleich zu den vorhandenen Sprachattributen sein.

3. NLP und Embeddings

**Word Embeddings:**

Verwendet Word2Vec oder GloVe, um Wörter in Vektoren umzuwandeln.

Damit könnt ihr ähnliche Wörter finden oder die semantische Bedeutung von Texten analysieren.

**Sentiment-Analyse:**

Bestimmt die Stimmung (positiv, negativ, neutral) der Liedtexte.

Gibt es Unterschiede in der Stimmung zwischen verschiedenen Genres oder Künstlern?

**Themenmodellierung (Topic Modeling):**

Findet heraus, welche Themen in den Liedtexten vorkommen.

Verwendet dafür zum Beispiel Latent Dirichlet Allocation (LDA).

Lokale Verarbeitung (Laptops)

**Datenreduktion:**

Da der Datensatz sehr gross ist, könnt ihr eine Stichprobe verwenden, um eure Modelle zu entwickeln.

Später könnt ihr sie auf dem gesamten Datensatz anwenden.

Visualisierungen

1. Häufigkeitsverteilungen

**Balkendiagramme:**

Häufigste Sprachen im Datensatz.

Top-Künstler nach Anzahl der Lieder oder Aufrufe.

Häufigste Genres (Tags).

Häufigste Wörter in Liedtexten (getrennt nach Sprachen oder Genres).

**Histogramme:**

Verteilung der Liedlängen (Wortanzahl).

Verteilung der Aufrufe (Views).

Verteilung der Erscheinungsjahre.

2. Beziehungen und Zusammenhänge

**Streudiagramme:**

Zusammenhang zwischen Liedlänge und Aufrufen.

Zusammenhang zwischen Erscheinungsjahr und Aufrufen.

**Heatmaps:**

Korrelation zwischen verschiedenen numerischen Attributen (z. B. Aufrufe, Erscheinungsjahr).

Häufigkeit von Wortkombinationen in verschiedenen Genres.

**Liniendiagramme:**

Entwicklung der Beliebtheit von Genres über die Jahre.

3. Textvisualisierungen

**Wortwolken:**

Häufigste Wörter in Liedtexten (für den gesamten Datensatz oder nach Genres/Sprachen).

**Balkendiagramme mit Wortfrequenzen:**

Top 10 oder Top 20 der häufigsten Wörter in Liedtexten.

4. Geografische Visualisierungen (falls relevant)

**Karten:**

Verteilung der Künstler nach Ländern (falls ihr diese Information habt).

Wichtige Hinweise