Universität Leipzig
Fakultät für Mathematik und Informatik
Institut für Informatik

Exposé
-
Genre-specific topics in song lyrics

Eine Projektarbeit von Niko Konzack, Benjamin Friedl und Lucas Zetzsche

Leipzig, Januar, 2024

Betreuender Hochschullehrer:
Dr. Andreas Niekler
Computional Humanities, Institut für Informatik

Einleitung

Lassen sich anhand der Themen von Musik Lyrics dessen Genre erkennen? Bisher gibt es dazu keine Forschung. Verwandte Arbeiten lassen sich in 4 Kategorien einteilen. Erstens gibt es ein wenig Forschung zur Klassifikation von Musikgenres anhand von Lyrics. Allerdings werden hier keine Topic Models benutzt [1,2]. Zweitens wurde die Qualität der Zuordnung von Themen zu Lyrics erforscht [3,4,5]. Der Zusammenhang mit Musikgenres wurde dabei aber nicht beleuchtet. Drittens wurde schon versucht, anhand der Themen von Lyrics das Sentiment zu klassifizieren [6,7]. Dies stellt einen ähnlichen Prozess im Vergleich zur Klassifikation von Genres anhand der Themen der Lyrics dar. Die technische Umsetzung der Forschenden wirft allerdings Fragen auf. Generelle Methoden zur Klassifikation mittels oder mit Hilfe von Topic Modellen wurden allerdings auch schon untersucht [8,9].
Dies führt uns zu unserer Forschungsfrage: Können wir verschiedene Musikgenres durch die Themen, die den Lyrics zugrunde liegen, unterscheiden? Dies soll anhand der durch Topic Modellen entstehenden Themen und der Performance eines darauf bauenden Klassifikators untersucht werden.

Daten

Als Textkorpus für das Training des Topic Modells werden Songtexte über die API von Genius [10] gesammelt. Genius ist eine Online-Wissensdatenbank, die es Nutzern ermöglicht, Songtexte hochzuladen, zu bearbeiten und mit Anmerkungen zu annotieren. Zudem können Liedtexte mit Tags versehen werden, was es ermöglicht, ihnen Genres zuzuordnen. Genius unterscheidet zwischen den fünf Hauptgenres Country, Pop, Rap, R&B und Rock und über 1000 weiteren Subgenres. Nach Möglichkeit sollen bis zu 1000 der populärsten Songs der fünf Hauptgenres für das Modell verwendet werden.
Für eine größere Liste an Genres besteht die Möglichkeit, eine spätere Analyse mit weiteren durch das Free Music Archive [11] vorgeschlagenen Genres durchzuführen. In diesem Fall würden zusätzlich zu den oben genannten fünf Genres die Genres Blues, Jazz, Experimental, Folk, Classical, Electronic zur Analyse hinzukommen.
Andere durch das Free Music Archive vorgeschlagene Top-Level Genres wie beispielsweise “Historic” werden in diesem Fall nicht mit aufgeführt, da Genius diese Genres nicht mit dieser Bezeichnung aufführt.

Methoden

Die Methoden lassen sich in zwei Teile aufgliedern. Zuerst werden die durch ein Topic Modell entstehenden Themen manuell evaluiert. Als Topic Modell hat sich hierfür das LDA als nützlich erwiesen. Zur Wiedererkennung von Themen spezifischer Genres wird für jedes Genre auch ein eigenes Topic Modell trainiert. Spezielles Augenmerk gilt außerdem auch dem Einfluss der Anzahl an Topics und der Festlegung anderer Hyperparameter. Die Fähigkeit des Topic Modells, unterschiedliche Genres zu trennen, lässt sich allerdings auch anhand der Performance eines, auf den Topic Verteilungen trainierten Klassifikators erkennen. Dies dient uns anschließend als quantitative Analyse und wird eventuell für verschiedene Versionen des Topic Modells durchgeführt.

Ergebnisse

Die Anwendung des Topic-Modells (LDA) auf die Songtexte wird es uns ermöglichen, verschiedene Themen zu identifizieren, die den Lyrics zugrunde liegen. Durch die manuelle Evaluation werden wir Zusammenhänge zwischen diesen Themen und den definierten Musikgenres feststellen. Wir werden auch mögliche nutzlose Topics erkennen und analysieren, um die Qualität der entstehenden Topics zu bewerten. Die Topics für jedes Genre werden extrahiert und daraufhin untersucht, ob sie in den Gesamt-Topics wiedererkennbar sind. Wir werden auch die Variation der Anzahl der Topics und die Festlegung anderer Hyperparameter berücksichtigen, um die Robustheit des Modells zu testen.
Im Kontext der Vorbereitung der Features für den Klassifikator werden wir spezifische Topic-Modelle für jedes Genre trainieren. Die Hyperparameter werden sorgfältig festgelegt, und verschiedene Versionen des Topic-Modells werden getestet. Die Auswirkungen auf die Performance des Klassifikators werden als quantitative Analyse dienen, um die Fähigkeit des Topic-Modells zur Trennung verschiedener Genres zu bewerten.
Die Ergebnisse, also die identifizierten Themen und die Performance des Klassifikators, werden dabei mit Plots dargestellt und anhand dessen erläutert. Verwendete Skripte und Datensätze werden dabei in einem GitHub Repository hinterlegt und an entsprechenden Stellen auch verlinkt.

Diskussion

Die Interpretation der erwarteten Ergebnisse wird es uns ermöglichen, die Forschungsfrage bezüglich der Unterscheidbarkeit verschiedener Musikgenres anhand der identifizierten Themen zu beantworten. Die wichtigsten Erkenntnisse werden zusammengefasst, und potenzielle Einschränkungen des Ansatzes werden vorweggenommen. Die Diskussion wird die Bedeutung der erwarteten Ergebnisse hervorheben und Schlussbemerkungen zur Anwendbarkeit des gewählten Ansatzes in der musikbezogenen Klassifikation bieten. Eventuelle Verbesserungsmöglichkeiten und offene Fragen werden ebenfalls erörtert, um eine umfassende Reflexion über die geplante Forschungs-Durchführung zu gewährleisten. Zum Schluss wird noch ein Blick in die Zukunft gewagt und potentielle Arbeiten, die auf unseren Ergebnissen aufbauen könnten angesprochen.

Literaturverzeichnis

[1]: View of Data Science Approach to Compare the Lyrics of Popular Music Artists (uii.ac.id)
[2]: aclanthology.org/C14-1059.pdf
[3]: Assessing Quality of Unsupervised Topics in Song Lyrics | SpringerLink
[4]: lyric_topic_modeling.pdf (alenlukic.com)
[5]: Indonesians' Song Lyrics Topic Modelling Using Latent Dirichlet Allocation | IEEE Conference Publication | IEEE Xplore
[6]: Exploiting Topic Modelling to Classify Sentiment from Lyrics | SpringerLink
[7]: Music mood classification based on lyrical analysis of Hindi songs using Latent Dirichlet Allocation | IEEE Conference Publication | IEEE Xplore
[8]: Statistical topic models for multi-label document classification | Machine Learning (springer.com)

[9]: Using LDA Topic Models as a Classification Model Input | by Marc Kelechava | Towards Data Science
[10]: Genius API
[11]: FMA: A dataset for music analysis

Hilfsmittelverzeichnis:

- ChatGPT, Version 3.5

Redlichkeitserklärung:

Wir erklären, dass wir in einem Verzeichnis alle verwendeten Hilfsmittel (KI-Assistenzsysteme) deklariert und ihre Verwendung bei den entsprechenden Textstellen angegeben haben.