*Exposé Million Song Dataset – Klassifizierung und Recommender-System*

**Hintergrund / Motivation / Problemstellung**

Im Rahmen des Bachelor-Wirtschaftsinformatik Projektes zum Thema Data-Science für Web-Applikationen bei Dr. Kai Brüssau und Dr. Frank Schwartz entschieden sich die Verfasser dieser wissenschaftlichen Arbeit für die Erarbeitung eines Recommender-Systems für die Empfehlung von Musik. Hierzu beschäftigten sich die Verfasser mit der gegebenen praktischen Problemstellung aus dem wissenschaftlichen Bereich der Data Science und verwendeten den Datensatz Million Song Dataset.[[1]](#footnote-1) Die Motivation für dieses Thema begründete sich auf dem persönlichen Interesse an Recommender-Systemen und Musik.

Der ausgewählte Datensatz umfasst für jeden einzelnen Song der insgesamt eine Million Songs in der Datenbasis umfangreiche Metadaten, die uns Informationen zu dem jeweiligen Song liefern. Die vorliegenden Daten sind eine unstrukturiert Datenmenge, eine Einteilung bzw. Kategorisierung liegt in der Ausgangssituation nicht vor.

Zu dieser Ausgangslage hinsichtlich der vorhandenen Daten, kamen die geringen Kenntnisse des Projektteams zum Thema Data-Mining und hier insbesondere den unterschiedlichen statistischen Methoden, die zur Analyse verwendet werden können. Einen ersten Eindruck erhielt das Projektteam im Rahmen der einführenden Themen-Vorträge.

**Zielsetzung**

Das übergeordnete Ziel dieser Projektarbeit stellte für das Projektteam ein funktionierendes Recommender-System zur Empfehlung von Musik dar, basierend auf vom User präferierten, eingegebenen, ebenfalls im Datensatz enthaltenen, Songs.

Hierfür galt es zunächst, sich notwendige theoretische Grundlagen ausgewählter statistischer Methoden anzueignen und ein Verständnis dessen zu entwickeln. Darauf aufbauend konnte das Projektteam die für diese Problemstellung geeigneten Methoden auswählen und anwenden.

Entscheidend für den Erfolg der Projektarbeit war die erfolgreiche Klassifikation der in dem Datensatz enthaltenen Songs, über die zugehörigen ausgewählten Metadaten. Diese Klassifikation bildet die Grundlage des Recommender-Systems für Anwender.

Am Ende des Projektes sollte ein funktionierender Algorithmus zur bereits oben beschriebenen Song-Klassifikation über ausgewählte Metadaten und ein funktionierender Algorithmus zur Recommender-Funktion stehen. Als weiteres optionales Ziel entschied sich das Projektteam für eine graphische Benutzeroberfläche, eine UI, als Anwendung.

Abschließend sah die Zielsetzung eine schriftliche Aufarbeitung des erarbeiteten theoretischen Wissens und der praktischen Lösungsumsetzung der Problemstellung vor.

**Vorgehensweise / Methodik zur Lösung des Problems**

Zu Beginn des Projektes stand die Auswahl einer Statistik- bzw. Programmierumgebung zur Analyse der Metadaten. Das Projektteam entschied sich für Python, da hier auf bereits vorhandenen Programmierkenntnissen aufgebaut werden konnte und es sich für die Projektzwecke gut anbot. Des Weiteren musste die Einarbeitung und Einrichtung der Datenbank erfolgen. Somit waren die grundlegenden technischen Voraussetzungen für den Projektstart geschaffen.

Danach arbeitete sich das Projektteam in den Bereich der Data Science und das Data Mining ein, um den technischen Voraussetzungen die theoretischen hinzuzufügen.

Der nächste Schritt war das Einlesen aller Datensätze in die Datenbank und die Auswahl der geeigneten Metadaten bzw. Parameter aus dem Datensatz. Das Projektteam entschied sich zunächst für „loudness“, „hotttnesss“, „tempo“, „time signature“, „key“, „mode“. Mithilfe dieser Parameter und den statistischen Methoden erhoffte man sich eine gute Unterteilung der einzelnen Datensätze zu erreichen. Ob hier im Laufe des weiteren Projektes noch Anpassungen vorgenommen werden und zum Beispiel noch ein Parameter wegfällt, ist noch offen. Bei dem Parameter „hotttnesss“ fehlen leider bei knapp der Hälfte aller Songs die Werte, bei den anderen Parametern sind die Werte vollständig. Um diesen Aspekt zu behandeln führen wir eine Imputation mithilfe von MICE durch.

Ein Bild, das Text, Screenshot enthält.

Mit hoher Zuverlässigkeit generierte Beschreibung

In der Folge mussten wir die Daten mittels Mean Normalization normalisieren, um die unterschiedlichen Wertebereiche der einzelnen Parameter einem Standard anzupassen. Diese wurden dann gespeichert und geclustert.

Während dieses Clusterns erhält jeder Song ein Label, in welchem Cluster er sich befindet. Gibt ein User nun einen favorisierten Song ein, werden von unseren Algorithmen andere Songs nach ihrer Distanz zu den eingegebenen Songs analysiert. Songs, die sich in demselben Cluster-Label befinden wie der vom User eingegebene Song erhält dann die höchsten Punktzahlen. Die Songs mit den höchsten Punktzahlen soll der User dann als empfohlene Songs ausgegeben bekommen.

Aktuell sind die Ergebnisse unseres Clusterings noch nicht zufriedenstellend. Die besten Silhouette\_Scores erhalten wir weiterhin für sehr geringe Cluster-Anzahlen. Dies kann noch nicht als optimales Ergebnis gewertet werden.

**Aufbau der Arbeit, insbesondere Gliederung mit Seitenumfang**

1. Einführung (2 Seiten)
   1. Problemstellung
   2. Zielsetzung
2. Theoretische Grundlagen
   1. Data-Mining Methoden (8 Seiten)  
      2.1.1. Clustering  
      2.1.2. Random Forest

2.1.3. Imputation

* 1. Musiktheorie (2 Seiten)

1. Methodik und Vorgehensweise (8 Seiten)

3.1. ...

3.2. ...

1. Ergebnis (2 Seiten)
2. Kritische Würdigung (2 Seiten)
3. Schluss (1 Seite)

**Zeitplan (Literaturrecherche, Implementierung, etc.)**

|  |  |  |  |
| --- | --- | --- | --- |
| # | Kategorie | Aktivität | Zeitplanung |
| I. | **Projektvorbereitung** | Einrichtung von Github-Projekt | x |
|  |  | Einrichtung von Overleaf-Projekt | x |
|  |  | Einrichtung von und Einarbeitung in Datenbank | x |
|  |  | Analyse und Beschreibung der Problemstellung | x |
|  |  |  |  |
|  |  |  |  |
| II. | **Recherche** | Recherche Data-Mining Methoden | 15.05. |
|  |  | |- Clustering |  |
|  |  | |- Random Forest |  |
|  |  | |- Imputation (missing values) |  |
|  |  |  |  |
| III. | **Implementierung** | Datensätze finden und in die Datenbank eintragen | 22.05. |
|  |  | Back end | 15.05. bis 22.05 |
|  |  | |- Implementierung der Empfehlungsmethoden |  |
|  |  | Visualisierung der ersten Ergebnisse | 22.05. |
|  |  | Front end | 19.06. |
|  |  | |- Einrichtung von GUI |  |
|  |  | Qualitätskontrolle | 26.06 |
|  |  | |- Vergleich verschiedener Data-Science-Methoden |  |
|  |  |  |  |
|  |  |  |  |
| IV. | **Abschlussarbeit** | Literaturrecherche | 19.06. |
|  |  | Analyse und Interpretation der Ergebnisse | ab 19.06. |
|  |  | Anfertigung der Abschlussarbeit | ab 19.06. |

**Erste Literaturübersicht**

Für die vorliegende Projektarbeit mit der Problemstellung eines Recommender-Systems für Musik sind als Literatur und wissenschaftliche Theorien in erster Linie Exemplare zu den ausgewählten Data-Mining Methoden sowie zur Musiktheorie, Musikpsychologie und Musikanalyse relevant. Hier befindet sich das Projektteam aktuell noch in der Recherche. Folgende Literatur wurde bisher bereits ausgewählt:

* Rose D. (2016) Using a Data Science Life Cycle. In: Data Science. Apress, Berkeley, CA
* Ghatak, A. (2017): Machine Learning with R, Springer, Singapore
* Auhagen, W., Bullerjahn, C. und von Georgi, R. (2015): Musikpsychologie - anwendungsorientierte Forschung, Hogrefe, Göttingen
* Auhagen, W., Bullerjahn, C. und Louven, C. (2017): Musikpsychologie - Akustik und musikalische Hörwahrnehmung, Hogrefe, Göttingen

1. https://labrosa.ee.columbia.edu/millionsong/ [↑](#footnote-ref-1)