Projektbericht: Spotify Analyse

1st Eric Kaufmann
Jena, Germany
eric.kaufmann@uni-jena.de

2nd Maria Gogolev
Jena, Germany
maria.gogolev@uni-jena.de

***Abstract*—In diesem Projektbericht wurden zwei Datenstze von Kaggle verwendet, verbunden und anschlieend analysiert. Dabei werden Merkmale von Songs durch die Einteilung in Playlisten betrachtet. Mittels Erstellung einer neuen Metrik knnen Eigenschaften populrer Songs bestimmt und Vernderungen im Laufe der Zeit verglichen werden.**

I. Datenstze

Als Datenstze wurden zwei von kaggle.com verwendet:

1) Spotify 1.2M+ Songs[1]
2) Spotify Playlists[2]

Ersteres besteht, wie der Name es schon verdeutlicht, aus ca. 1.2 Millionen (1,204,025 genau) unterschiedlichen Spotify Songs. Zu jedem Song sind dabei 24 Eigenschaften gegeben. Eine Auflistung aller Eigenschaften mit Beschreibung ist in Tabelle (I) zu finden. Der Datensatz liegt als CSV vor umfasst ungefhr 346MB. Die Daten wurden dabei mittels der offiziellen Spotify-API generiert. Jeder Song kann mittels einer eindeutigen ID beschrieben werden. Durch eine GET-Anfrage auf *https://api.spotify.com/v1/tracks/*$\{id\}$ knnen Informationen zum Album, Songtitel, Artisten und Publikationsdatum extrahiert werden. Mittels zweiter GET-Anfrage auf *https://api.spotify.com/v1/audio-features/*$\{id\}$ knnen die restlichen Eigenschaften, wie *danceability*, *energy* oder *liveness* bestimmt werden.

Der Datensatz ermglicht es Songs anhand von Eigenschaften zu sortieren. Beispielsweise kann man verschiedene Eigenschaften wie die *danceability*, *valence* oder *energy* der Songs ber die Zeit betrachten. Eine andere exemplarische Anwendungsmglichkeit ist ein Algorithmus zur Songempfehlung, bei dem man Songs mit hnlichen Eigenschaften versucht zu verbinden.

Um die Songs besser unterteilen zu knnen wurde in dieser Analyse ein zweiter Datensatz verwendet, der Spotify Playlist Datensatz. Dieser besteht aus ca. 2.8 Millionen unterschiedlichen Songs, die in ca. 162 Tausend Playlists von ungefhr 16 Tausend User eingeteilt sind. Trotz einer Gre von ca. 1.8 GB ist der Aufbau des Datensatzes recht einfach. Wie in Tabelle (II) erkennbar besitzt der Datensatz nur vier Spalten. Dabei hat ein User mit einer *user_id* eine Playlist mit der Playlistbezeichnung *playlistname* angelegt. Songs, bestimmt durch den *trackname* und *artistname*, knnen somit einer Playlist zugehrig sein. Selbstverstndlich knnen dabei Songs mehrfach vorkommen, wenn sie beispielsweise in unterschiedlichen Playlists enthalten sind oder in einer Playlist sogar mehrfach aufkommen.

[1]siehe https://www.kaggle.com/datasets/rodolfofigueroa/spotify-12m-songs
[2]siehe https://www.kaggle.com/datasets/andrewmvd/spotify-playlists

TABLE I
Beschreibung der Spalten des Spotify 1.2M+ Songs Datensatz

Eigenschaft	Beschreibung
id	Song-ID
name	Songtitel
album	Albumtitel
album_id	Album-ID
artists	Liste der Artisten
artists_ids	Liste der Artisten-IDs
track_number	Tracknummer des Songs im Album
disc_number	Albumnummer
explicit	Song ist explicit
danceability	Eignung des Songs zum tanzen ($\in [0, 1]$)
energy	Intensitt und Aktivitt des Songs ($\in [0, 1]$)
key	Tonart
loudness	Lautstrke (dB)
mode	Modus (Dur/Moll)
speechiness	Sprachanteil ($\in [0, 1]$)
acousticness	Akustik eines Songs ($\in [0, 1]$)
instrumentalness	Instrumentalanteil ($\in [0, 1]$)
liveness	Wahrscheinlichkeit einer Livebertragung ($\in [0, 1]$)
valence	Wertigkeit ($\in [0, 1]$, 0=negativ und 1=positiv)
tempo	Tempo (BPM)
duration_ms	Dauer (ms)
time_signature	Taktart
year	Verffentlichungsjahr
release_date	Verffentlichungsdatum (YYYY-MM-DD)

TABLE II
Beschreibung der Spalten des Spotify Playlist Datensatz

Eigenschaft	Beschreibung
user_id	User-ID eines Spotify Nutzers
artistname	Name des Artisten
trackname	Songtitel
playlistname	Playlistbezeichnung

Mgliche Anwendungsfelder dieser Playlist ist beispielsweise die Untersuchung der Popularitt einiger Songs durch die Anzahl der Aufkommen des Liedes in unterschiedlichen Playlisten. Mehr dazu spter.

Um die Songeigenschaften des ersten Datensatzes mit den Playlisteinteilungen des zweiten Datensatzes zu verbinden, mssen die beiden Datenstze verbunden werden. Dafr versuchen wir an den Playlist-Datensatz mittels left-join die Eigenschaften anzuheften. Dabei sollen die gemeinsamen Schlssel der Songtitel mit Artisten sein. Um dies zu erreichen mssen jedoch vorher die Daten vorbereitet werden. Dafr wurden zunchst die Spaltennamen *name* und *artists*, beim Spotify

1.2M+ Datensatz, bzw. *trackname* und *artistname*, beim Spotify Playlistdatensatz, vereinfacht und in *track* und *artist* umbenannt. Somit sind die beiden Schlssel gleich bezeichnet.

Der nchste Schritt ist es die Liste von Artisten, welche in jeder Zeile der *artist*-Spalte beim Spotify 1.2M+ Songs Datensatz zu finden ist, aufzusplitten. Diese haben nmlich eine folgende Form:

$$[\text{'artist}_1\text{'}, \text{'artist}_2\text{'}, \ldots, \text{'artist}_n\text{'}].$$

Ziel ist es, dass die Zeile des Datensatzes n-Mal repliziert wird und in jeder neuen Zeile nur der jeweilige artist_i steht. Zum Schluss wurden die Datentypen aller Spalten angepasst und unrealistische Werte (z.B. Verffentlichungsjahr 0) aussortiert.

Nach dieser Vorbereitung der Daten ist es nun mglich den left-join des Spotify 1.2M+ Songs Datensatz an den Spotify Playlist Datensatz ber die gemeinsame Schlssel *track* und *artist* durchzufhren. Zur Vereinfachung der spteren Analyse wurde anschlieend alle Zeilen mit *NA*-Werten aus dem verbundenen Datensatz entfert. Somit werden auch alle Zeilen entfert, bei dem der Spotify Playlist Datensatz keine Werte fr Songs aus dem Spotify 1.2M+ Songs Datensatz findet.

Die Verbindung der Datenstze hat jedoch einen groen Datenverlust zur Folge. Nach der Vorverarbeitung der einzelnen Datenstze sind beim Spotify 1.2M+ Songs Datensatz noch ca. 1,194,000 und beim Spotify Playlist Datensatz noch ca. 2,795,000 unterschiedliche Songs brig. Nach dem join sind jedoch nur noch ca. 210,000 unterschiedliche Songs brig. Die Grnde dafr sind recht unterschiedlich. Zum einen gibt es natrlich Songs, die in dem Spotify Playlist Datensatz vorkommen, jedoch nicht im Spotify 1.2M+ Songs Datensatz. Ein anderer Grund sind die Differenzen in der Bezeichnung der Werte der gemeinsame Schlssel *track* und *artist*. Auf Spotify gibt es sehr hufig Songs, welche zwar einen bestimmen Artisten haben, jedoch ein anderer User dieser Song hochgeladen hat. Somit existiert zwar dieser Song, der Schlssel *artist* stimmt aber nicht berein. Generell werden zu jedem bekannten Song viele Unterschiedliche Versionen von vielen Nutzern hochgeladen. So gibt es von einem Song eine offizielle Version, ein Radio-Edit, eine Live-Version und verschiedene Remix. Auch Features mit anderen Artisten bringen hufig Probleme. Somit ist es mglich, dass ein Song im einen Datensatz die Form *"song_titel"* hat und im anderen *"song_titel (feat. artist2)"*. Im Spotify 1.2M+ Songs Datensatz sind die einzelnen Artisten als Liste gespeichert. Beim Spotify Playlist Datensatz sind diese zusammen als String gespeichert. Daraus entsteht das Problem, dass die Artisten im Playlist Datensatz die Form *"artist1, artist2 and artist3"* bzw. *"artist1 & artist2"* haben knnen. Eine Teilung an Kommata ist aber auch nicht mglich, da es Artisten gibt, welche Kommata in ihrem offiziellen Namen haben. Somit wird die Schnittmenge der beiden Schlsselmengen noch kleiner.

II. Analyse und Interpretation

A. Entwicklung der Musik ber die Zeit

Aus den vorliegenden Daten geht hervor, dass sich die Musik im Laufe der Jahre in Bezug auf verschiedene Attribute deutlich verndert hat, und gewisse Trends ersichtlich sind. Die Analyse bezieht sich auf Abbildung 1. Im Folgenden werden nicht alle Attribute aufgegriffen, da bei ihnen Statistiken wie Standardabweichung und Durchschnittswert weniger aussagekrftig sind (bspw "Key" also der Notenschlssel).

- Danceability Beginnend mit der danceability sehen wir, dass der Minimal- und Maximalwert ber die Jahre leicht auseinander gehen. Nach einem Absinken der danceability in den 1940 Jahren zeigt der Durchschnittswert wieder einen leichten Aufwrtstrend. Die Standardabweichung bleibt relativ stabil ber die Zeit.
- Energy Auch bei der Energie ist ein hnlicher Trend zu beobachten: Der Minimalwert bleibt relativ konstant, der Maximalwert steigt stetig an, der Durchschnittswert zeigt einen Aufwrtstrend und die Standardabweichung nimmt im Laufe der Zeit zu. Dies deutet darauf hin, dass die Musik im Laufe der Jahre energiegeladener geworden ist, es gleichzeitig aber mehr Diversitt in diesem Attribut gibt.
- Loudness Bei der Lautstrke ist ein sinkender Minimalwert und eine leichre Zunahme des Maximalwerts und Durchshcnittswertes zu verzeichnen. Musik wird also leicht lauter.
- Mode Der Durchschnittswert des Modus bleibt ber die gesamte Zeit berhalb von 0.5, was aussagt, dass stets mehr Musik in Dur produziert wird als in Moll, wobei ein leichter Abstieg des Durchschnittswertes in den letzten jahren zu verzeichnen ist. Es wurde also in den letzten Jahren mehr in Moll produziert.
- Speechiness Das Minimum des Sprachanteils bleibt nahe bei 0, da Instrumentalmusik frher wie auch heute relevant ist. Der anfangs geringe Maximalwert entwickelt sich hingegen nach oben. Durchschnittswert und Standardabweichung bleiben gering was darauf hindeutet das der Groteil der Musik nach wie vor instrumental ist. Ansonsten zeigt die Sprachlichkeit keinen starken Trend.
- Acousticness Die Akustizitt zeigt ber die Jahre einen Rckgang der Durchschnittswerte, was darauf hindeutet, dass die Musik im Laufe der Jahre elektronischer geworden ist. Die Standardabweichung fr das Attribut "Akustizitt" nimmt im Laufe der Zeit zu, was darauf hindeutet, dass die Verteilung der Akustizittswerte im Laufe der Zeit immer vielfltiger und unterschiedlicher wird. Elektronische Musik ersetzt akustische Musik also nicht, sondern kommt zustzlich hinzu.
- instrumentalness Bei der Instrumentalitt ist ein auf und ab zu sehen, ohne klaren Trend hinsichtlich wachsendem oder fallendem Instrumentalanteil. Die Standardabweichung bleibt konsistent hoch, was bedeutet, dass nach wie vor viele Lieder mit hohem als auch niedrigem Instrumentalanteil produziert werden.

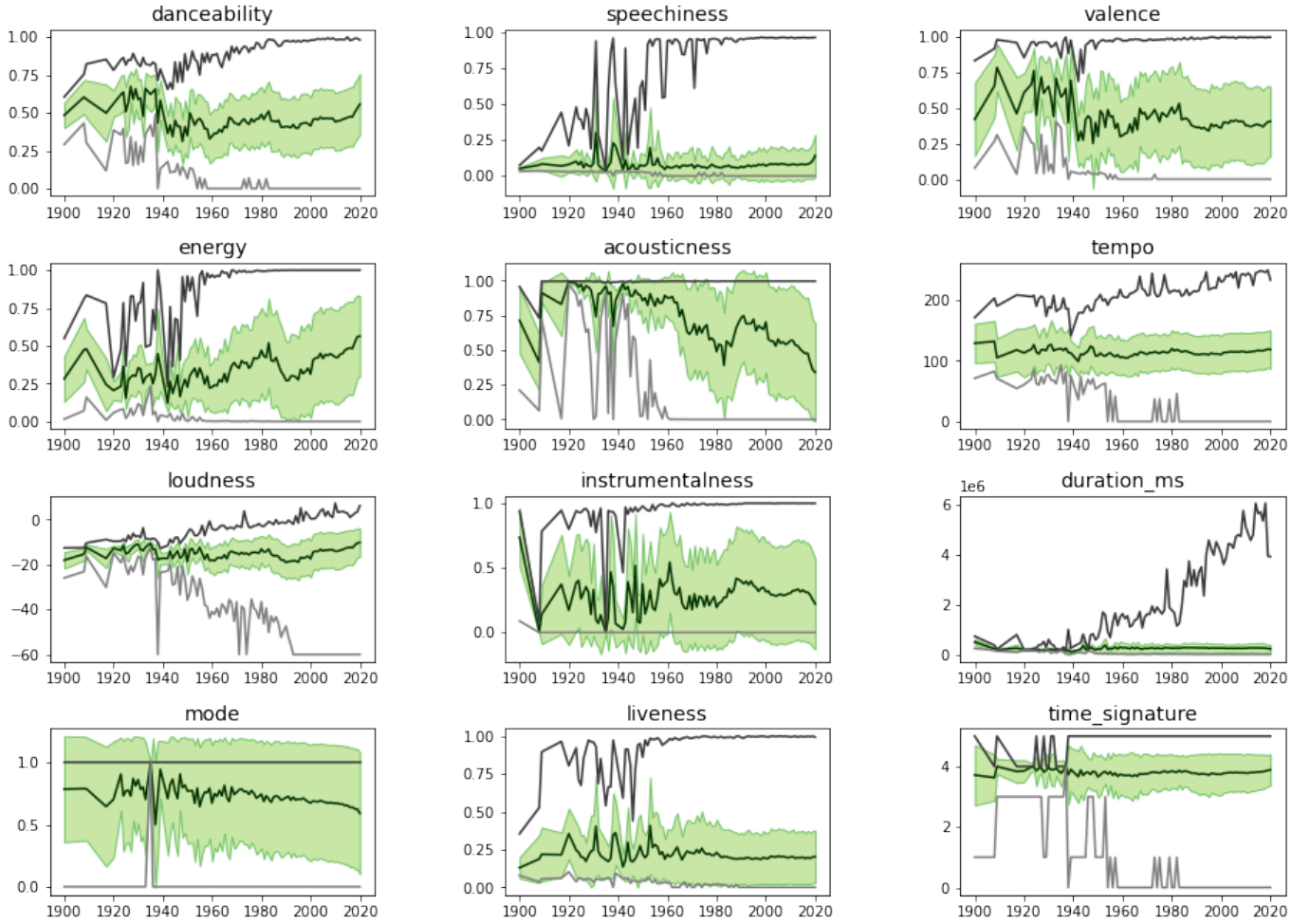

Fig. 1. Entwicklung der Musik-Attribute von 1900 bis 2020. Die in der Mitte verlaufende Linie ist der Durchschnittswert, die obere Linie das Maximum, die untere Linie das Minimum und der grne Bereich ist die Standardabweichung im gegebenen Jahr.

- liveness Auch ”Liveness” zeigt keinen eindeutigen Trend. Der Durchschnittswert ist relativ niedrig mit einer geringen Standardabweichung, was darauf hindeutet dass der Groteil der Musik nicht live bertragen wurde.
- valence Die Valenz zeigt keinen konstanten Abstieg der Statistiken, aber einen Reduktion des Durchschnittswerts ab den 1940ern hnlich zur ”Danceability”, was darauf hindeutet, dass die Musik seit dieser Zeit weniger positiv ist.
- Tempo Das Maximaltempo wird ber die Jahre hher und das Minimaltempo wird niedriger whrend der Durchschnittswert relativ konstant im mittleren Bereich der Attributskala bleibt. Die Standardabweichung ist niedrig und konstant. Die meisten Lieder sind also weder langsam noch schnell, aber es wird ber die Jahre immer wieder mit neuen Tempo-Rekorden experimentiert.
- Duration Die Dauer zeigt einen starken Anstieg des Maximalwertes. Minimal- und Durchschnittswerte bleiben relativ konsistent und gering, genauso wie die Standardabweichung. Die Musik ist in Bezug auf die Dauer eher stabil und vorhersehbar.
- Time signature Der Anteil der Songs mit 4/4-Takt ist im Laufe der Zeit relativ stabil geblieben, die meisten Songs sind im 4/4-Takt, die Standardabweichung der Taktart ist gering und im Laufe der Zeit stabil, was darauf hindeutet, dass die meisten Songs, nach wie vor, hauptschlich 4 Schlge pro Takt haben.

Insgesamt erweckt die Grafik den Anschein, dass die Musik im Laufe der Jahre energiegeladener und etwas tanzbarer und lauter geworden ist, wobei die durchschnittliche Tanzbarkeit und das Energieniveau allmhlich gestiegen sind. Die durchschnittliche Akustik und Instrumentalitt schwanken. Die Musik scheint aber insgesamt zu mehr elektronischen oder synthetischen Klngen zu tendieren, da sich der Anteil der akustischen Musik ber die Jahre eindeutig verringert. Fr die Attribute deren Durchschnittswert einem klaren Trend folgt, also der relative Rckgang akustischer Elemente und der Anstieg der Energie, lsst sich auch ein Anstieg in der Standardabweichung, also der Diversitt in der Ausprgung des Attributs feststellen.

B. Popularitt

Das Ziel in diesem Abschnitt besteht darin, die Statistiken der Attribute durchschnittlicher Lieder und Knstler mit den Statistiken populrer Lieder Knstler zu vergleichen, um herauszufinden was Popularitt ausmacht. In keinem der beiden Datenstzen gibt es Informationen ber die Beliebtheit, weshalb wir zunchst ein Beliebtwert fr jeden Knstler und fr jeden Song wie folgt ausrechnen: Es wird gezhlt wie oft das gegebene Lied/ der gegebene Knstler in Playlisten aufgenommen wurde. Das gibt dann einen Schtzwert fr die Popularitt. In Tabelle III sieht man die 10 Lieder die am hufigsten in Playlisten vorkommen, und Tabelle IV zeigt die top 10 Knstler die am hufigsten in Playlisten aufgenommen wurden.

track	artist	count
Ho Hey	The Lumineers	1547
Kids	MGMT	1319
Rather Be (feat. Jess Glynne)	Clean Bandit	1279
Do I Wanna Know?	Arctic Monkeys	1204
Chandelier	Sia	1169
Don't Stop Believin'	Journey	1114
Sail	AWOLNATION	1110
Take Me Out	Franz Ferdinand	1084
Creep	Radiohead	1066
Summer	Calvin Harris	1058

TABLE III
TOP 100 POPULRSTE LIEDER

artist	count
Beyonc	3476
Radiohead	3406
Arctic Monkeys	3038
Coldplay	2855
Michael Jackson	2593
MGMT	2516
Bob Dylan	2307
Foo Fighters	2212
The Strokes	2189
Bruce Springsteen	2131

TABLE IV
TOP 100 POPULRSTE KNSTLER

1) Vergleich populrer Lieder mit durchschnittlichen Liedern: In Abbildung 2 wird deutlich, dass die Durchschnittswerte der populren Songs wesentlich niedriger fr "Danceability", fr die Lautstrke, fr akustische Elemente, fr Tempo und fr die Taktart. Das heit populre Lieder eignen sich im Durchschnitt weniger zum Tanzen, sind viel leiser, langsamer, und haben einen geringeren akustischen Anteil. Sie haben weniger Schlge pro Takt. Populre Songs sind jedoch energetischer. Auerdem kann man erkennen, dass die Standardabweichung fr die top 100 beliebtesten Lieder in fast allen Attributen geringer ist, was darauf hindeutet, dass sich populre Lieder untereinander in ihren Attributen hnlicher sind. Man kann auf Grund dieser Resultate populre Lieder (Lieder die oft in Playlisten auf Spotify aufgenommen werden) wie folgt charakterisieren: sie haben trotz geringerem Tempo und geringer Lautstrke viel Energie leicht positiver, und generell weniger Variation in den Attributen. Eventuell eignen sie sich also eher zum Hren im Hintergrund.

2) Vergleich populrer Knstler mit durchschnittlichen Knstlern: Zunchst haben wir die Durchschnittswerte fr jeden Knstler und jedes Attribut durch Aggregation aller Lieder des Knstlers ber das arithmetische Mittel errechnet. Dannach wurden die top 100 beliebtesten Knstler gefiltert, und mit allen Knstlern insgesamt verglichen.

In Abbildung 3 sieht man viele hnlichkeiten zu den Ergebnissen in Abbildung 2.

Ein paar kleinere Unterschiede sind die im Durchschnitt geringere "Valence" der populren Knstler im Vergleich zum Durchschnittsknstler und die hnlchkeit der durchschnittlichen Energielevel. Die kleine Standardabweichung ist auch hier wieder indikativ fr die hnlichkeit der populren Knstler untereinander. Was man in beiden Abbildungen erkennen kann, ist dass Popularitt mit weniger Eignung zum Tanzen, langsamerem Tempo, weniger Schlgen pro Takt, geringerer Lautstrke, weniger Instrumentalanteil, weniger akustischen Elementen und weniger Varianz in den Attributen einhergeht.

3) Vergleich der Eigenschaften populrer Lieder mit den Liedern populrer Knstler: In Abbildung 4 sieht man die Statistiken beliebter Knstler gegenbergestellt zu den Statistiken beliebter Lieder. Man kann folgende Unterschiede feststellen: Die Varianz der Attribute scheint bei beliebten Knstlern noch strker eingeschrnkt zu sein als bei den beliebtesten Songs, was bedeutet dass sich beliebte Knstler untereinander mehr hneln als beliebte Lieder. Weiterhin haben beliebte Knstler etwas weniger energetische Lieder, aber dafr mit hherem Akustik-Anteil.

REFERENCES

[1] Pichl, Martin; Zangerle, Eva; Specht, Gnther: "Towards a Context-Aware Music Recommendation Approach: What is Hidden in the Playlist Name?" in 15th IEEE International Conference on Data Mining Workshops (ICDM 2015), pp. 1360-1365, IEEE, Atlantic City, 2015.

[2] Figueroa, Rodolfo. (2020). Spotify 12M Songs [Dataset]. Kaggle. https://www.kaggle.com/datasets/rodolfofigueroa/spotify-12m-songs

[3] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[4] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[5] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[6] K. Elissa, "Title of paper if known," unpublished.

[7] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[8] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[9] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

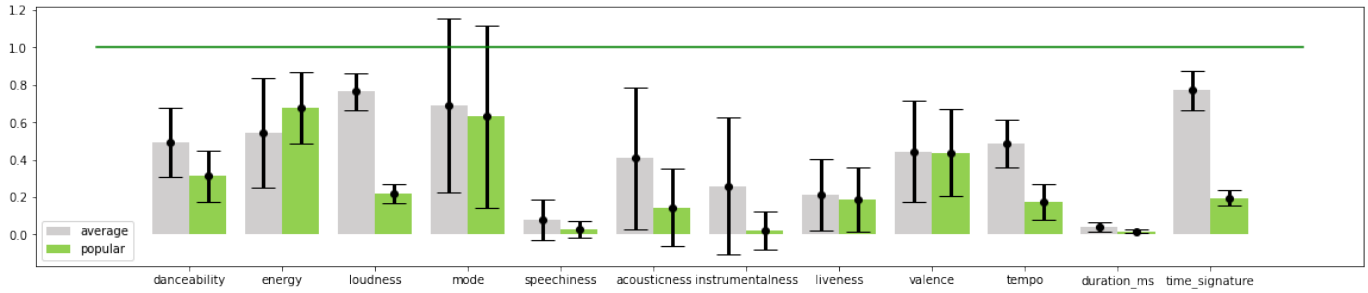

Fig. 2. Vergleich des Durchschnittswerts und der Standardabweichungen verschiedener Attribute der den Top 100 beliebtesten Songs mit allen Songs. Rechts (grn) sind die populren Lieder dargestellt, und links(grau) sieht man die durchschnittlichen Lieder. Da die Attribute unterschiedliche Wertebereiche haben, werden die Durchschnittswerte und die Standardabweichungen jeweils mit dem globalen Maximum und dem globalen Minimum des Attributs skaliert.

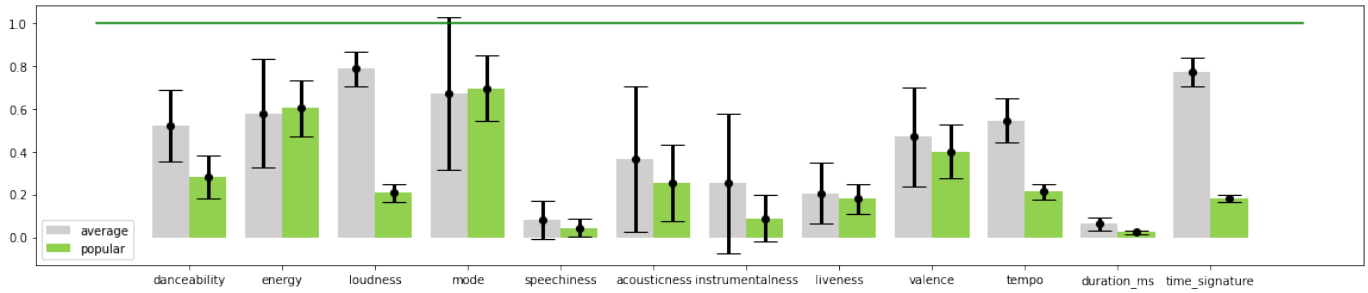

Fig. 3. Vergleich des Durchschnittswerts und der Standardabweichungen verschiedener Attribute der den Top 100 beliebtesten Knstler mit allen Knstlern insgesamt. Rechts (grn) sind die populren Knstler dargestellt, und links(grau) sieht man die durchschnittlichen Artisten. Da die Attribute unterschiedliche Wertebereiche haben, werden die Durchschnittswerte und die Standardabweichungen jeweils mit dem globalen Maximum und dem globalen Minimum des Attributs skaliert.

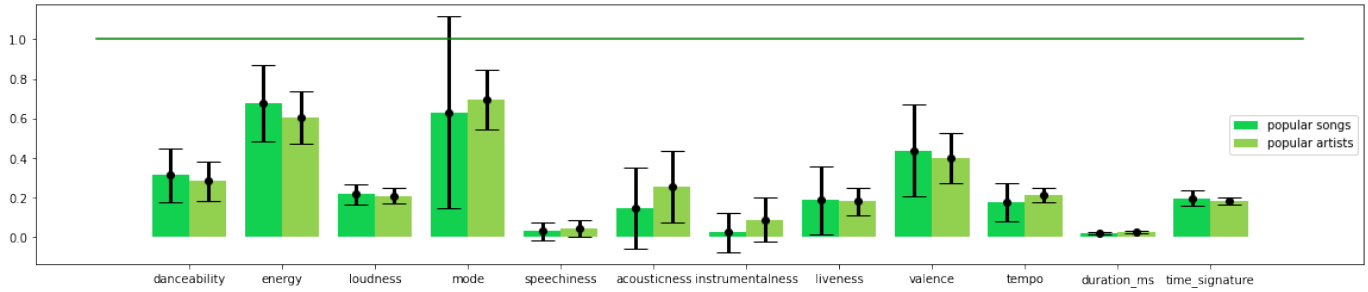

Fig. 4. Vergleich des Durchschnittswerts und der Standardabweichungen verschiedener Attribute der Durchschnittssongs der Top 100 beliebtesten Knstler mit den top 100 beliebtesten Songs insgesamt. Rechts (grn) sind die populren Knstler dargestellt, und links sieht man die populren Songs. Da die Attribute unterschiedliche Wertebereiche haben, werden die Durchschnittswerte und die Standardabweichungen jeweils mit dem globalen Maximum und dem globalen Minimum des Attributs skaliert.