Beskrivelse af slutprojekt

Gruppe 28

Navn	Studienr
Jacob Odgaard Hausted	201510912
Kasper Gnutzmann Andersen	201607263
Gill Lumer-Klabbers	201607384

Dataset

Gruppen har valgt at arbejde med et datasæt baseret på Spotify's API, "*Spotify Tracks DB*"[1]. Datasættet indeholder data fra sange som findes på spotify platformen, og hver sang er i datasættet beskrevet med 16 features. Nogle af disse features knytter sig sangens identifikation og generelle attributter, som id, navn, genre og artist. Disse er hentet fra sangens 'Track' [2]API objekt. Resten er sangens 'Audio features' [3], som beskriver forskellige aspekter af sangens *lyd* og *feel.*

Herunder fremgår alle datasættets features, opdelt efter deres tilhørende API objekt:

Identifikation og generelle features (*fra 'Track'-objekt*):

- Genre
- Artist name
- Track name
- Track ID
- Popularity

***Lyd* og *feel (fra 'Audio features'-objekt)*:**

- Acousticness
- Danceability
- Duration in ms
- Energy
- Instrumentalness
- Key
- Liveness
- Loudness
- Mode
- Speechiness
- Tempo
- Time signature
- Valence

Datasættet er fundet på kaggle og indeholder data for ~ 233 tusinde sange, ~ 10000 pr genre (26 unikke genre inkluderet i sættet).

Det bør påpeges at datasættet kan udbygges, da spotify's API står frit tilgængeligt, hvorfor datasættet er tiltænkt som et startpunkt for gruppen til at arbejde på. Siden datasættet effektivt er en delmængde af spotify's fulde database, findes der rig mulighed for at tilpasse datasættet hvis mængden ikke findes tilstrækkelig til den givne opgave.

Udover de tidligere benævnte features kan listen udvides, da API'et stiller et yderligere *Audio Analysis object* [4] til rådighed. Dette objekt indeholder resultater fra den bagvedliggende digitale signal analyse af de individuelle sange – effektivt betydende at hver feature nu findes repræsenteret ved en "*timbre"* vektor. Derudover vil der være mulighed for at opdele være sang i flere segmenter og få beskrevet hvert segments

[1] https://www.kaggle.com/zaheenhamidani/ultimate-spotify-tracks-db
[2] https://developer.spotify.com/documentation/web-api/reference/tracks/get-track/
[3] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/
[4] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-analysis/

features (ovenstående for det givne segment) med tilhørende confidence-niveauer – altså et mål for hvor præcise featuresne er for det givne segment.

Anvendelse:

Hovedidéen for gruppens anvendelse af datasættet vil være at prøve at prædiktere en feature ud fra de andre, evt. genre og popularitet ud fra featuresne defineret i *lyd* og *feel.* Til netop dette ønsker gruppen også at udvide datasættet med *Audio Analysis* objektet, hvortil der vil blive mulighed for at teste sammenhængen imellem spotify's confidence mål og egne fundne type I og II fejl. Datasættet ligger i høj grad op til supervised learning, men der er også muligheder for unsupervised learning, f.eks clustering af audiofeatures.

Gruppen har derudover leget med idéen om at udvælge en delmængde af datasættet og tilføje en ny feature, "*likeability",* som vil være et mål for hvor godt et givent gruppemedlem kan lide sangen. Dette vil være i forhåbningen om at der i sidste ende vil kunne laves en algoritme der vil være i stand til at prædiktere hvilke sange et givent gruppemedlem kan lide (findes der en sammenhæng mellem de oplyste features og medlemmets musiksmag). Problematikken her vil dog i høj grad være tiden det vil tage at manuelt skulle label en stor nok delmængde til at kunne træne modellen.