MASTER INFO
Data mining
Projet

Julien Blanchard

Février 2019

Consignes

- **Le travail est à réaliser en binôme.**
- **Les noms des étudiants binômés sont à envoyer à julien.blanchard@univ-nantes.fr avant le 15 mars 2019.**
- **Toute oeuvre de plagiat entre binômes sera sanctionnée par une note nulle (plagieurs comme plagiés).**
- **Toute oeuvre de plagiat depuis internet sera sanctionnée par une note nulle.**
- **Date limite de remise des travaux (sources+rapport) sur Madoc dans la section "Data Mining (X2II020)" : 5 mai 2019**

L'utilisation massive de playlists (listes de lecture) fait partie des nouveaux usages apportés par les services de streaming musical. La composition des playlists les plus populaires revêt un enjeu majeur du marché contemporain de la musique, d'abord parce que les playlists permettent d'engranger des écoutes et donc des royalties, mais aussi car elles sont un véritable moyen de promotion permettant à un titre ou un artiste de gagner en notoriété et de rencontrer son public.

Parmi les offres de streaming musical, Spotify est un service fondé en 2008 en Suède. Faisant partie des plateformes de musique en ligne les plus utilisées, il génère pour les artistes une part toujours croissante de leurs revenus. Dans ce projet, nous intéressons à cinq playlists Spotify très suivies : Love Pop, Exception Française, Electronic Circus, Metal Essentials, et Coffee Table Jazz.

1 Données

Les données sont issues du scraping du site Spot on Track qui historise les données de Spotify (charts et playlists). Elles sont constituées de deux fichiers :

- le fichier `playlists.data` qui décrit la composition des cinq playlists (instantané hebdomadaire depuis mai 2017) ;
- le fichier `tracks.data` qui décrit les chansons qui apparaissent dans les playlists à l'aide de 10 variables calculées par Spotify :
  - Le `BPM` (battements par minute) est une mesure exprimant le tempo de la musique ou le rythme cardiaque.
  - La tonalité (`Key`) – voir ici la correspondance entre notes latines et notes anglo-saxonnes.
  - Le `Mode`, mineur ou majeur.

— `Danceability` describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0% is least danceable and 100% is most danceable.
— `Valence` is a measure from 0% to 100% describing the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
— `Energy` is a measure from 0% to 100% and represents a perceptual measure of intensity and activity. Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.
— `Acousticness` is a confidence measure from 0% to 100% of whether the track is acoustic. 100% represents high confidence the track is acoustic.
— `Instrumentalness` predicts whether a track contains no vocals. 'Ooh' and 'aah' sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly 'vocal'. The closer the instrumentalness value is to 100%, the greater likelihood the track contains no vocal content. Values above 50% are intended to represent instrumental tracks, but confidence is higher as the value approaches 100%.
— `Liveness` detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live. A value above 80% provides strong likelihood that the track is live.
— `Speechiness` detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 100% the attribute value. Values above 66% describe tracks that are probably made entirely of spoken words. Values between 33% and 66% describe tracks that may contain both music and speech, either in sections or layered, including such cases as rap music. Values below 33% most likely represent music and other non-speech-like tracks.

La variable `url` sert à identifier chaque chanson de manière unique. Dans le fichier `playlists.data`, la variable `artists` peut contenir plusieurs noms d'artistes ou de groupes, avec des séparateurs variés (vous pourrez choisir de normaliser cette partie du modèle de données si besoin).

2 Travail demandé

2.1 Préparation des données

Les chansons en premières positions des playlists sont également les plus jouées. Des études ont montré que les playlists sont majoritaiement écoutées dans l'ordre, et que le nombre d'auditeurs d'une playlist décroît en fonction de la durée d'écoute avec un décrochage accéléré autour de la 12e chanson. Dès lors, il est plus intéressant pour un artiste de se retrouver en tête de playlist plutôt qu'à la fin.

Pour chaque chanson dans une playlist, calculez les nouvelles variables suivantes :

— la position-pic (meilleure position) obtenue par la chanson dans la playlist,
— un indicateur binaire qui vaut 1 si la chanson a une position-pic inférieure à 15,
— la durée pendant laquelle la chanson est apparue dans la playlist (en nombre de semaines),
— la position moyenne de la chanson dans la playlist (uniquement parmi les semaines où la chanson participe effectivement à la playlist),
— un indicateur binaire qui vaut 1 si la chanson a une position moyenne inférieure à 15.

Vous pouvez également traiter les problèmes de caractères accentués dans les noms d'artistes et titres de chansons.

2.2 Analyse exploratoire

- Pour la playlist de votre choix ou bien la totalité des playlists, réalisez une analyse statistique descriptive 1D et 2D des chansons décrites par leurs 10 variables (histogrammes, boxplots, nuages de points...). Commentez les résultats remarquables.
- Déterminez pour chaque playlist le profil de la chanson moyenne sur les 10 variables, et identifiez quelques morceaux qui s'en rapprochent le plus possible.
- Déterminez pour chaque playlist la chanson (ou l'artiste) qui est la mieux classée d'après les données, au sens de la position moyenne.
- Pour une chanson et la playlist de votre choix, visualisez l'évolution temporelle de la position.

2.3 Analyse exploratoire multidimensionnelle

- Pour la totalité des playlists, appliquez une méthode de réduction de dimension (ACP, t-SNE, MDS...) sur les chansons décrites par leurs 10 variables, puis projetez les chansons dans un plan. Interprétez. Isolez quelques chansons atypiques (*outliers*) pour chaque playlist.

2.4 Analyse prédictive

- En considérant toutes les playlists, ou bien en opposant deux playlists seulement, construisez un modèle permettant d'expliquer (=reconnaître) la playlist en fonction des 10 variables. Évaluez le modèle et interprétez.
- Pour la playlist de votre choix ou bien la totalité des playlists, construisez un modèle permettant d'expliquer la position[1] d'une chanson. Évaluez le modèle et interprétez.
- Pour la playlist de votre choix ou bien la totalité des playlists, construisez un modèle de scoring permettant de prédire la position[2] d'une chanson. Évaluez le modèle à l'aide d'une courbe ROC ou lift.

3 A remettre sur Madoc

- Vos sources : programmes et scripts, fichiers de projets si vous avez utilisé un logiciel dédié.
- Un rapport PDF (10 à 20 pages) qui présente le travail réalisé, avec en particulier :
  - les éventuels pré-traitements supplémentaires réalisés,
  - les algorithmes utilisés et hyperparamètres choisis,
  - l'évaluation et l'interprétation des résultats.

Vous pouvez aussi mélanger sources et rapport en déposant un notebook.

1. Au choix l'une des cinq nouvelles variables créées en préparation des données.
2. Au choix l'un des deux indicateurs binaires créés en préparation des données.