Recommending Music and the Audioscrobbler Data Set

Big Data Analytics

Steven Liatti & Jeremy Favre

Sommaire

- Introduction
- Présentation des données
- Nettoyage des données
- Statistiques
  - Statistiques étendues
- Features utilisées
- Algorithmes appliqués
- Tentative avec les graphes
- Conclusion

Introduction

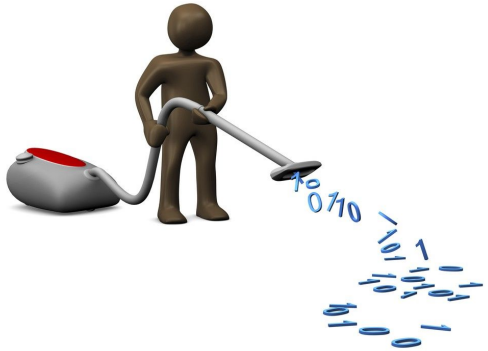
- Set de données fourni
  - Données d'une plateforme musicale
- Utilisateurs de la plateforme et les artistes qu'ils écoutent
  - Nombre d'écoutes par artiste
- Informations sur les noms corrects des artistes
- Objectifs
  - Exploiter ces données afin d'établir un modèle de recommandations musical
  - Recommendations d'artistes
  - Établir des statistiques
  - Appliquer des algorithmes de clustering
  - Utiliser la librairie GraphX de Spark

Présentation des données

user	artist	count
1000002	1	55
1000002	1000006	33
1000002	1000007	8
1000002	1000009	144

- Data set contient 3 fichiers
- user_artist_data.txt
  - Ligne: userId | artistId | count
- artist_data.txt
  - Ligne: artistId | name
- artist_alias.txt
  - Ligne: artistId | correctArtistId

Nettoyage des données



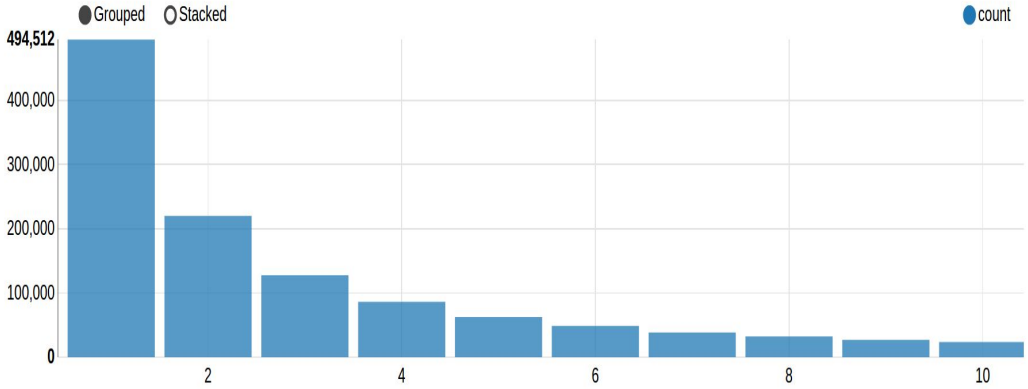
- Filtrage des “fakes”
  - Count pas cohérent
- Correction des noms mal orthographiés
- Suppression des artistes inconnus

Statistiques (1)

- Utilisateurs
  - Nombre
  - Nombre d’écoutes total par utilisateur (1er à 400'000)
  - Utilisateurs avec écoutes par artiste (“My Chemical Romance” 156’000)
  - Moyennes d’écoutes (162, max à 6734)

Statistiques (2)

- Artistes
  - Nombre
  - Top 20
  - Pire 20 -> pas trop de sens
  - Artistes écoutés moins de 10 fois
  - Médiane des écoutes



Statistiques (3)

- Nom d’artistes mal orthographiés
  - Nombre
  - “Pire” nom (Metallica)
  - En réalité ce ne sont pas des mauvaises orthographes

Plus de statistiques avec MusicBrainz (1)

- Data set pauvre
- API limitée
- Export CSV des tables de MusicBrainz
- Pour chaque nom d’artiste récupération des informations
  - Nom, type, sexe, genre, date de départ, date de fin, lieu géographique
- Joins

Plus de statistiques avec MusicBrainz (2)

- Types
- Actifs/inactifs
- Sexe
- Combien d'artistes féminines anglaises sont ou ont été actives depuis 1986 ?
- Provenance (USA)
- Artistes qui ont duré le plus (pas concluant)
- Tags (= genres)
- Jointure avec données initiales ?

Tentative avec les graphes

- GraphX
  - artistes et users comme noeud
  - count comme arêtes
- Malheureusement, pas fonctionné
  - Erreur non explicite

Description des features utilisées

- ALS
  - user
  - artist
  - count
- K-means & Gaussian Mixture Model
  - begin_date_year
  - 90% des échantillons utilisés pour l'entraînement
  - 10% des échantillons utilisés pour le test

Algorithmes appliqués

- ALS
  - Recommandation de musiques
- Clustering
  - K-means
    - k=8 à faire varier pour obtenir des meilleures performances
  - Gaussian Mixture Model

Conclusion

- Structure parallèle de Spark très utile sur notre data set
- Erreurs de Spark pas claires
- Data set pauvre

- Améliorations
  - Optimiser les hyperparamètres pour le recommandeur
  - Autre algorithme pour le recommandeur
  - Recommandations en temps réel