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
- Tentative avec les graphes
- Algorithmes appliqués
- Features utilisées
- Conclusion

Introduction

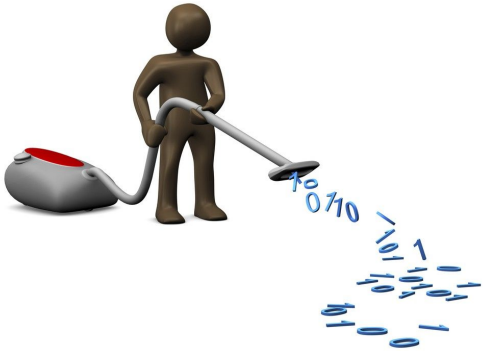
- Set de données fourni
  - Données d'une plateforme musicale
- Objectifs
  - Recommendations d'artistes
  - Établir des statistiques
  - Appliquer des algorithmes de clustering
  - Utiliser la librairie GraphX de Spark

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 (24 mio de lignes)
  - Ligne: userId | artistId | count
- artist_data.txt (1.8 mio de lignes)
  - Ligne: artistId | name
- artist_alias.txt (193’027 lignes)
  - Ligne: artistId | correctArtistId

Nettoyage des données



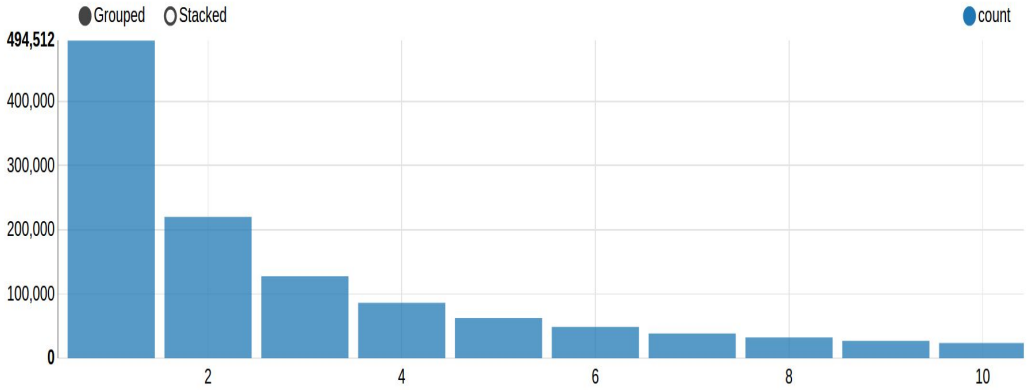
- Filtrage des “fakes”
  - Count pas cohérent
- Correction des noms mal orthographiés
- Suppression des artistes inconnus

Statistiques (1)

- Utilisateurs
  - Nombre (148’077)
  - 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 (1.6 mio)
  - Top 20 (Beatles, Metallica, Muse)
  - Pire 20 -> pas trop de sens
  - Artistes écoutés moins de 10 fois
  - Médiane des écoutes



Statistiques (3)

- Nom d’artistes mal orthographiés
  - Nombre (22’478, ~1.5%)
  - “Pire” nom (Metallica)
  - En réalité ce ne sont pas des mauvaises orthographes

Plus de statistiques avec MusicBrainz (1)

- Data set pauvre
- Import CSV dans Spark des tables de MusicBrainz
- Pour chaque nom d'artiste récupération des informations
  - Nom, type, sexe, genre, date de départ, date de fin, lieu géographique

Plus de statistiques avec MusicBrainz (2)

- Types (groupes et artistes seuls)
- Actifs/inactifs
- Sexe (3.5 * hommes)
- Provenance (USA)
- Artistes qui ont duré le plus (pas concluant)
- Combien d'artistes féminines anglaises sont actives depuis 1986 ? (28)
- Jointure avec données initiales ?

Tentative avec les graphes

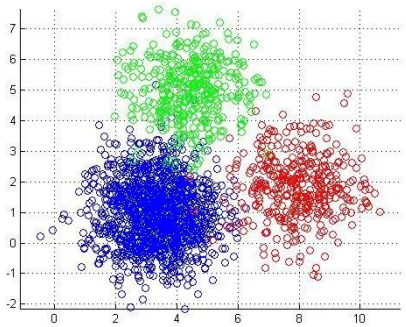
- GraphX
  - artistes et users comme noeuds
  - count comme arêtes
- Malheureusement, pas fonctionné
  - Erreur non explicite

Algorithmes appliqués

- Alternating Least Squares (ALS)
  - Recommandation de musiques
  - Area Under the Curve (AUC)
- Clustering
  - K-means
    - k=8 à faire varier pour obtenir des meilleures performances
  - Gaussian Mixture Model
    - K=5

id	name
1180	David Gray
378	Blackalicious
813	Jurassic 5
1255340	The Saw Doctors
942	Xzibit

id	name
2814	50 Cent
4605	Snoop Dogg
1037970	Kanye West
1001819	2Pac
1300642	The Game



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

Conclusion

- Data set, cleaning, statistiques, algorithmes
  - Bon tour des fonctionnalités proposées par Spark
- Structure parallèle de Spark très utile sur notre data set
- Data set pauvre
- Erreurs de Spark pas claires
- Améliorations
  - Optimiser les hyperparamètres pour le recommandeur
  - Autres algorithmes pour le recommandeur
  - Recommandations en temps réel (Oryx 2)