Hes·so Master

MSE | MASTER OF SCIENCE IN ENGINEERING

# Recommending Music and the Audioscrobbler Data Set

Projet BDA | 19.06.2020

Ayrton Dumas, Marco Mattei, Samuel Torche

# Introduction

Recommandation de musique

Utilisation de ALS : “Collaborative filtering”

# Données

Artistes

```
10584546 Nislije
10584550 ONEYA BASSIVITYMIXTAPE
10584556 Grant Green / Osunlade
10584564 Jae Kwon
6953654  Gwen vs. Britney
```

Ecoutes utilisateurs

```
1000002 1000007 8
1000002 1000009 144
1000002 1000010 314
1000002 1000013 8
1000002 1000014 42
```

Alias des artistes

```
10088054 1042317
1195917  1042317
1112006  1000557
1187350  1294511
1116694  1327092
```

# Questions d’analyse

1. Peut-on obtenir des résultats intéressants en utilisant des techniques de Market Basket Analysis telles que les règles d'association ?

2. Peut-on déduire les genres des musiques à l'aide d'un clustering des utilisateurs et leurs écoutes ?

3. Est-il possible de recommander non pas des artistes mais d'autres utilisateurs qui partagent les même goût avec ALS ?
   => Est-il possible de recommander des utilisateurs qui ont des goûts similaires à un certain utilisateur en utilisant du clustering ?

# MBA : Algorithme et optimisations

Preprocessing

```
+-------+--------------------+
|   user|               items|
+-------+--------------------+
|1000002|[1, 18, 28, 30, 5...|
|1000019|[1000010, 1000028...|
+-------+--------------------+
```

FPGrowth : spark.ml.fpm.FPGrowth vs spark.mllib.fpm.FPGrowth

Tuning

$$Support = \frac{frq(X,Y)}{N} \qquad Confidence = \frac{frq(X,Y)}{frq(X)}$$

# MBA - Evaluation

Trop de données -> subset arbitraire

*“However, mining association rules often results in a very large number of found rules, leaving the analyst with the task to go through all the rules and discover interesting ones.”*

1277 règles

```
[1275996,4267,979,0.7412831241283124]
R.E.M., Green Day,  -> Radiohead
[1275996,1274,976,0.737450462351387]
R.E.M., Red Hot Chili Peppers,  -> Nirvana
[1275996,1177,1205,0.7180659915060438]
R.E.M., Coldplay,  -> U2
[1275996,976,1274,0.7150176112712135]
R.E.M., Nirvana,  -> Red Hot Chili Peppers
```

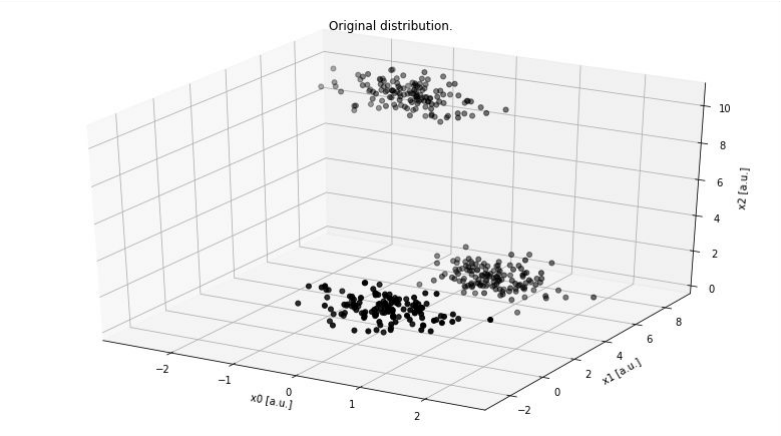
# Clustering des Genres - KMeans

Top 3 des artistes de chaque utilisateur

**(1005235, 1004983, 1239653)**

the constantine, the dismemberment plan et

Q and Not U (rock / post-hardcore)

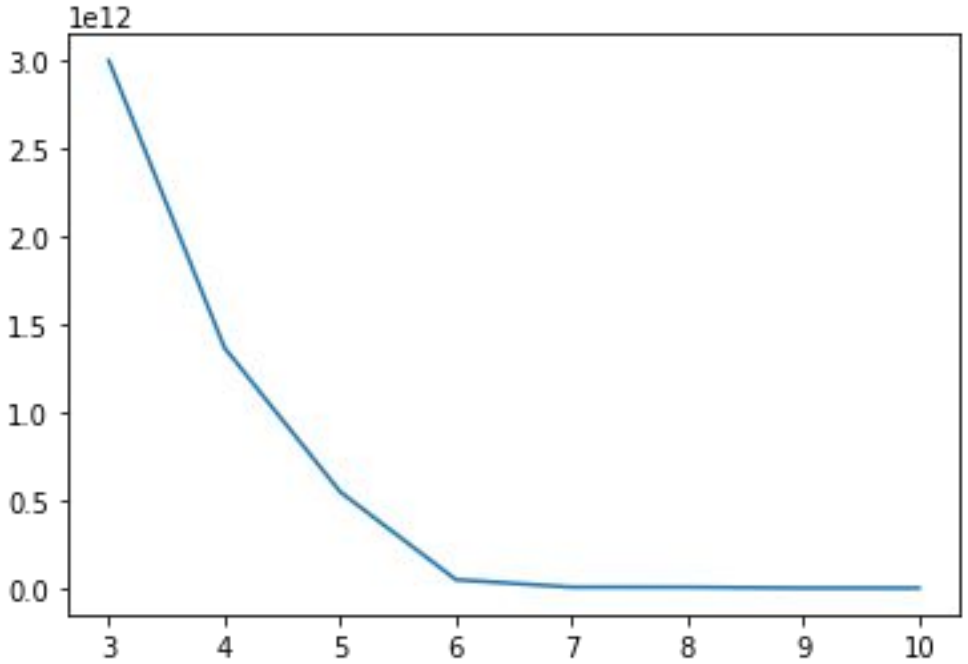


# Méthode du coude

K = 2, 3, 4, … 10

pour chaque K : Fit + MSE

6 clusters -> 6 genres de musique



# Clustering des genres - Evaluation

Assigner les centroids aux utilisateurs

Vérifier que les utilisateurs du même cluster écoute le même genre de musique

6 Genres à vérifier

# Recommandation d’utilisateurs - ALS

Suggérer à un utilisateur A **des utilisateurs qui écoutent des artistes similaires**

ALS pas adapté à ce genre de problème

ALS fonctionne sur le principe “user-item”

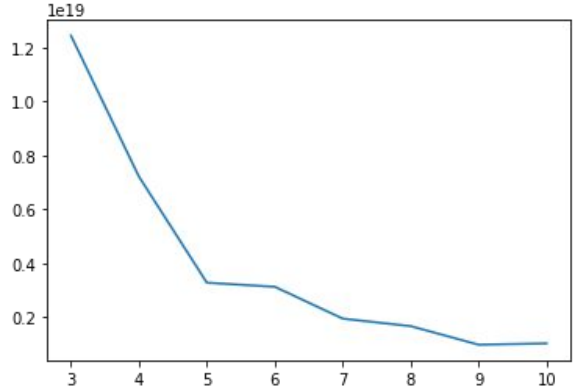
Ici => “user-user”

# Recommandation d’utilisateurs - Clustering

Clustering de toutes les données:

- X : Utilisateur
- Y : Artiste
- Z : Nb d’écoutes

Résultat -> Clusters d’utilisateurs ayant des goûts similaires

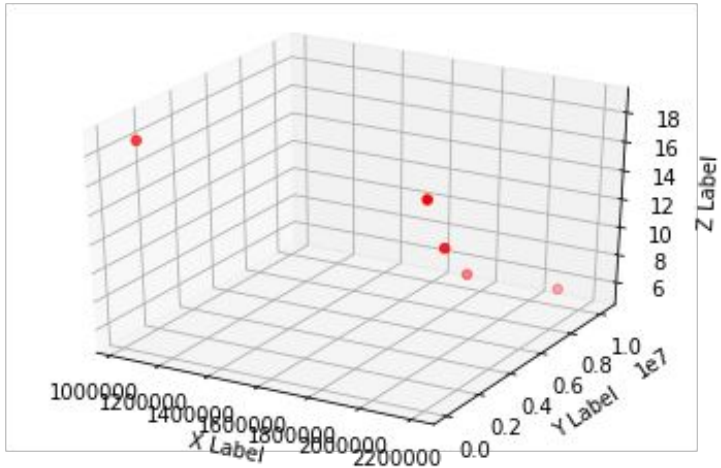


# Recommandation d’utilisateurs - Evaluation

5 groupes d’utilisateurs

Parcourir les données et assigner un centroïde à chacun

Pertinence à vérifier



# Améliorations futures

Nettoyage du dataset

MBA: plus de règles, plus de données

Evaluation des résultats de clustering, prendre plus d’artistes

Evaluation des résultats de clustering d’utilisateurs

Recommandation d’utilisateurs plus précise

# Conclusion

Proof of concept

Zeppelin n’était pas la bonne approche, trop lent pour du ML

Lien GitHub: https://github.com/samueltorche/bda-grpB-audio-recommender