30Music
Finding similarities between songs

Clément BAFFOS
Holly HEALEY
Victor LALEUW
Ping'an YANG

SOMMAIRE

01

Présentation de 30Music

Entities

- Albums
- Playlist
- Tags
- Tracks
- Users

Relations

- Events
- Love
- Sessions

01

Extraction des données

Format **.idomaar** : utilisé pour représenter des entités et leurs relations

- “users”

```
1   user    1   1116715959  {"lastfm_username":"000123","gender":"f","age":24,"country":"US","playcount":221012,"playlists":2,"subscribertype":"base"}
2   user    2   1163123792  {"lastfm_username":"000333","gender":"m","age":39,"country":"CZ","playcount":217535,"playlists":9,"subscribertype":"base"}
3   user    3   1184426573  {"lastfm_username":"00elen","gender":"f","age":,"country":"","playcount":49733,"playlists":2,"subscribertype":"base"}
4   user    4   1123157597  {"lastfm_username":"00Eraser00","gender":"m","age":32,"country":"DE","playcount":168054,"playlists":2,"subscribertype":"base"}
5   user    5   1171302116  {"lastfm_username":"00fieldsy","gender":"m","age":23,"country":"UK","playcount":45700,"playlists":2,"subscribertype":"base"}
```

- “love”

```
1   preference  1   -1 {"value":"love"} {"subjects":[{"type":"user","id":44542}], "objects":[{"type":"track","id":2785601}]}
2   preference  2   -1 {"value":"love"} {"subjects":[{"type":"user","id":44542}], "objects":[{"type":"track","id":2785590}]}
3   preference  3   -1 {"value":"love"} {"subjects":[{"type":"user","id":44542}], "objects":[{"type":"track","id":143076}]}
4   preference  4   -1 {"value":"love"} {"subjects":[{"type":"user","id":44542}], "objects":[{"type":"track","id":143037}]}
5   preference  5   -1 {"value":"love"} {"subjects":[{"type":"user","id":44542}], "objects":[{"type":"track","id":143052}]}
```

→ Nécessité de parser chaque fichier

01

Data cleaning

Nettoyage préliminaire des données :

- suppression des colonnes inutiles issues du parsing des données
- suppression des lignes avec des données manquantes ou aberrantes (0, NaN ou null)

	age	id	country	gender	lastfm_username	playcount	playlists	subscribertype
0	24	1116715959	US	f	000123	221012	2	base
1	39	1163123792	CZ	m	000333	217535	9	base
2	32	1123157597	DE	m	00Eraser00	168054	2	base
3	23	1171302116	UK	m	00fieldsy	45700	2	base
4	48	1273363051	UK	m	01dela	3869	7	base
5	24	1229981595	UK	m	01srainey	77775	4	base

Table "user" nettoyée

01

Data cleaning

Nettoyage en vue du partitioning :

- Transformation de certaines données qualitatives en données quantitatives pour pouvoir les exploiter

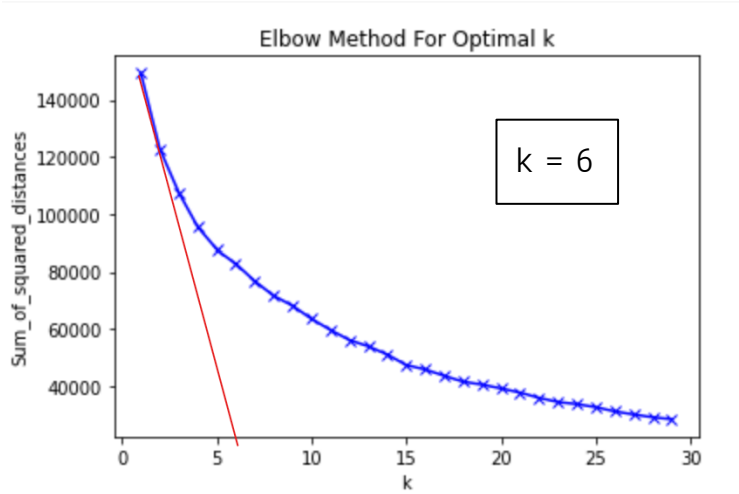
Variable qualitative	Variable quantitative
pays d'origine	proportion d'utilisateurs pour chacun des 6 pays les plus représentés
genre	proportion de femmes
âge	proportion d'utilisateurs par classe d'âge

SOMMAIRE

02

Partitioning : K-means

- Préparation de la donnée :
  - variables quantitatives uniquement
  - création d'un dictionnaire track-id/track_title pour interpréter les clusters
- Nombre de clusters déterminé par la méthode du coude



cluster_0

track_id	track_name
2748152	Tara+Putra/_/Dubland+Coastline
3732410	Kalabi/_/Last+Words
1316489	Gil+Scott-Heron/_/Evolution+(And+Flashback)
1419055	Jacco+Gardner/_/Clear+the+Air
843899	Deacon+Blue/_/Fergus+Sings+The+Blues
1155214	Faithless/_/Postcards
1674649	Lyambiko/_/Feeling+Good
476434	Boyd+Rice+and+Fiends/_/Watery+Leviathan
3015283	The+Zombies/_/If+It+Don%27t+Work+Out
2087924	Orchestral+Manoeuvres+in+the+Dark/_/The+Misund...

02

Analyse des clusters obtenus

cluster_0	cluster_1	cluster_2
• Musiques principalement écoutées en **Allemagne** (proportion moyenne d'Allemands sur ces chansons de 90.8%) • Pas de classe d'âge spécifique	• Musiques principalement écoutées aux **Etats-Unis** par la tranche d'âge **31-40** (proportions supérieures à 70% pour ces 2 critères dans ce cluster)	• Musiques principalement écoutées aux **Etats-Unis** par la tranche d'âge **21-30** (proportions supérieures à 88% pour ces 2 critères dans ce cluster)

02

Analyse des clusters obtenus

cluster_3

- Musiques principalement écoutées par la tranche d'âge **41-60** (proportions supérieures à 90%)
- Pas de pays spécifique mais légère prédominance aux **Etats-Unis** (59%)

cluster_4

- Musiques principalement écoutées par la tranche d'âge **21-30**
- Musiques écoutées dans tous les pays de façon relativement uniforme

cluster_5

- Musiques principalement écoutées au **Royaume-Uni** par les tranches d'âge **21-30** et **31-40**

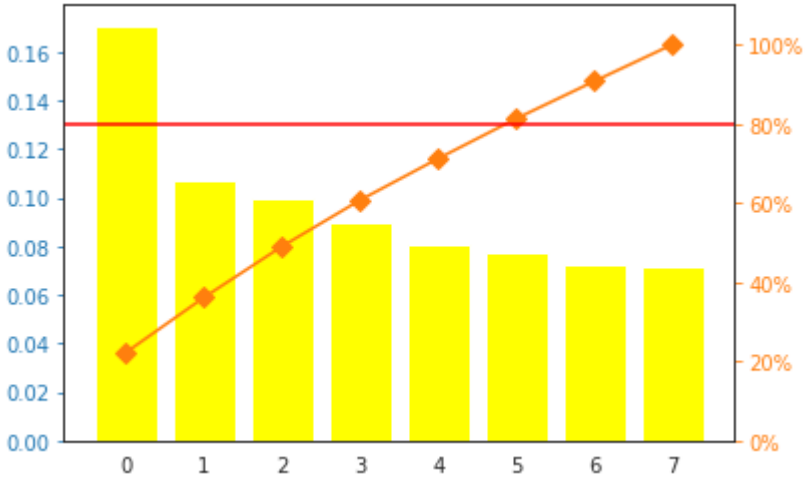
SOMMAIRE

03

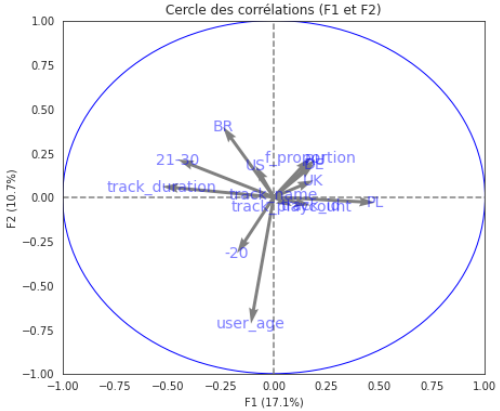
Analyse en composante principale

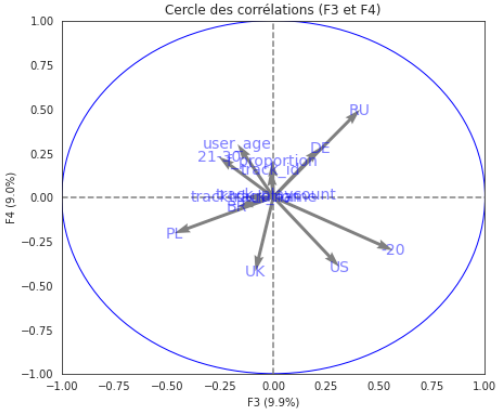
- Les méthodes density-based ne permettent pas de traiter un grand nombre de dimension

Part de variabilité expliqué par les composantes



On conserve 4 composantes qui expliquent 60% de la variance



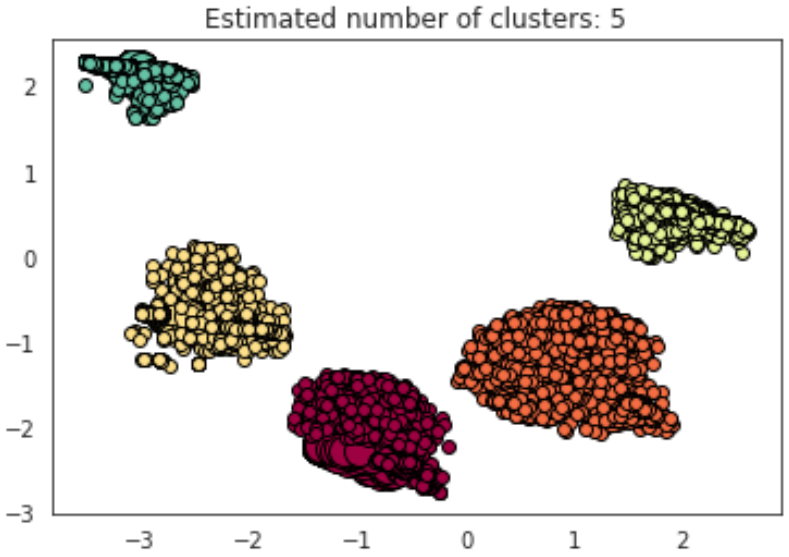


03

Density-based : DBSCAN

Méthode DBSCAN :

- eps = 0,5
- min_samples = 5000



- Nombre de clusters : 5
- Proportion de noise points : 52%

cluster_0

	track_id	track_name
0	107	009+Sound+System/_/Born+To+Be+Wasted
1	122	009+Sound+System/_/Number+Two
16	1378	10,000+Maniacs/_/City+of+Angels
28	1497	1000+Funerals/_/Igneous+Lips
29	1499	1000+Funerals/_/Night%27s+Dew+(Shape+of+Despai...
30	1503	1000+Funerals/_/Sutured+Lips
31	1505	1000+Funerals/_/Your+Fancy
43	2253	108/_/Deathbed
44	2255	108/_/Declarations+On+A+Grave
66	2667	10+Ft.+Ganja+Plant/_/Midnight+Landing

02

Analyse des clusters obtenus

cluster_0	cluster_1	cluster_2
• Musiques principalement écoutées aux **Etats-Unis** par la tranche d'âge **31-40** (proportions supérieures à 97% pour ces 2 critères dans ce cluster) • **21%** de femmes	• Musiques principalement écoutées aux **Etats-Unis** par la tranche d'âge **21-30** (proportions supérieures à 95% pour ces 2 critères dans ce cluster) • **31%** de femmes	• Musiques principalement écoutées aux **Etats-Unis** par la tranche d'âge **41-60** (proportions supérieures à 98% pour ces 2 critères dans ce cluster) • **18%** de femmes

02

Analyse des clusters obtenus

cluster_3	cluster_4
• Musiques principalement écoutées aux **Royaume-Uni** par la tranche d'âge **21-30** (proportions supérieures à 97% pour ces 2 critères dans ce cluster) • **29%** de femmes	• Musiques principalement écoutées en **Allemagne** par la tranche d'âge **41-60** (proportions supérieures à 99% pour ces 2 critères dans ce cluster) • **96%** d'hommes

SOMMAIRE

04

Traitement des données

Sélection des 500 musiques les plus aimées :

- Création de la matrice avec les musiques et les utilisateurs
  - Ligne : Musique
  - Colonne : Musique
  - Valeur : Le nombre d'utilisateurs qui aiment la Ligne et la Colonne en même temps

	14820	60007	71340	74094	74289	74350	74403	94083	119301	133352	...
14820	0.0	0.0	0.0	0.0	14.0	19.0	14.0	12.0	1921200.0	2.0	...
60007	0.0	0.0	0.0	0.0	14.0	10.0	9.0	7.0	0.0	230.0	...
71340	0.0	0.0	0.0	0.0	1.0	1.0	7.0	0.0	0.0	2.0	...
74094	0.0	0.0	0.0	0.0	657.0	235.0	668.0	17.0	0.0	7.0	...
74289	14.0	14.0	1.0	657.0	0.0	1392.0	807.0	44.0	0.0	9.0	...
...	...	...	...	...	...	...	...	...	...	...	...

04

Spectral Clustering

	node_1	node_2	distance
0	60007	14820	100.0
1	71340	14820	100.0
2	71340	60007	100.0
3	74094	14820	100.0
4	74094	60007	100.0
...	...	...	...
124745	4770627	3755697	100.0
124746	4770627	3756296	100.0
124747	4770627	3770868	100.0
124748	4770627	3770874	100.0
124749	4770627	4768900	100.0

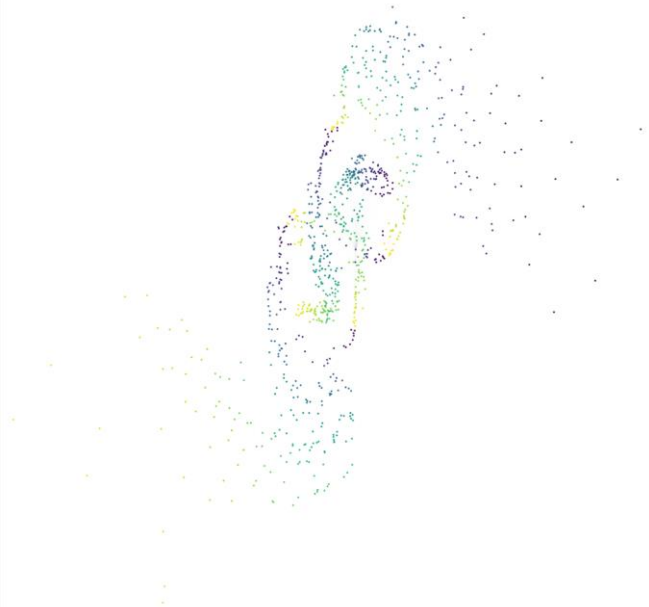
Préparation des données:

- Musique comme les nodes
- L'inverse des valeurs de matrice comme la distance entre deux nodes
- Pour les valeurs zéros on met 100 comme la distance

- *S'il y a des utilisateurs aiment les deux musiques, la distance est ≤ 1, sinon, c'est 100
- *Il n'y pas de duplications des combinaisons entre "node_1" et "node_2"

04

Spectral Clustering

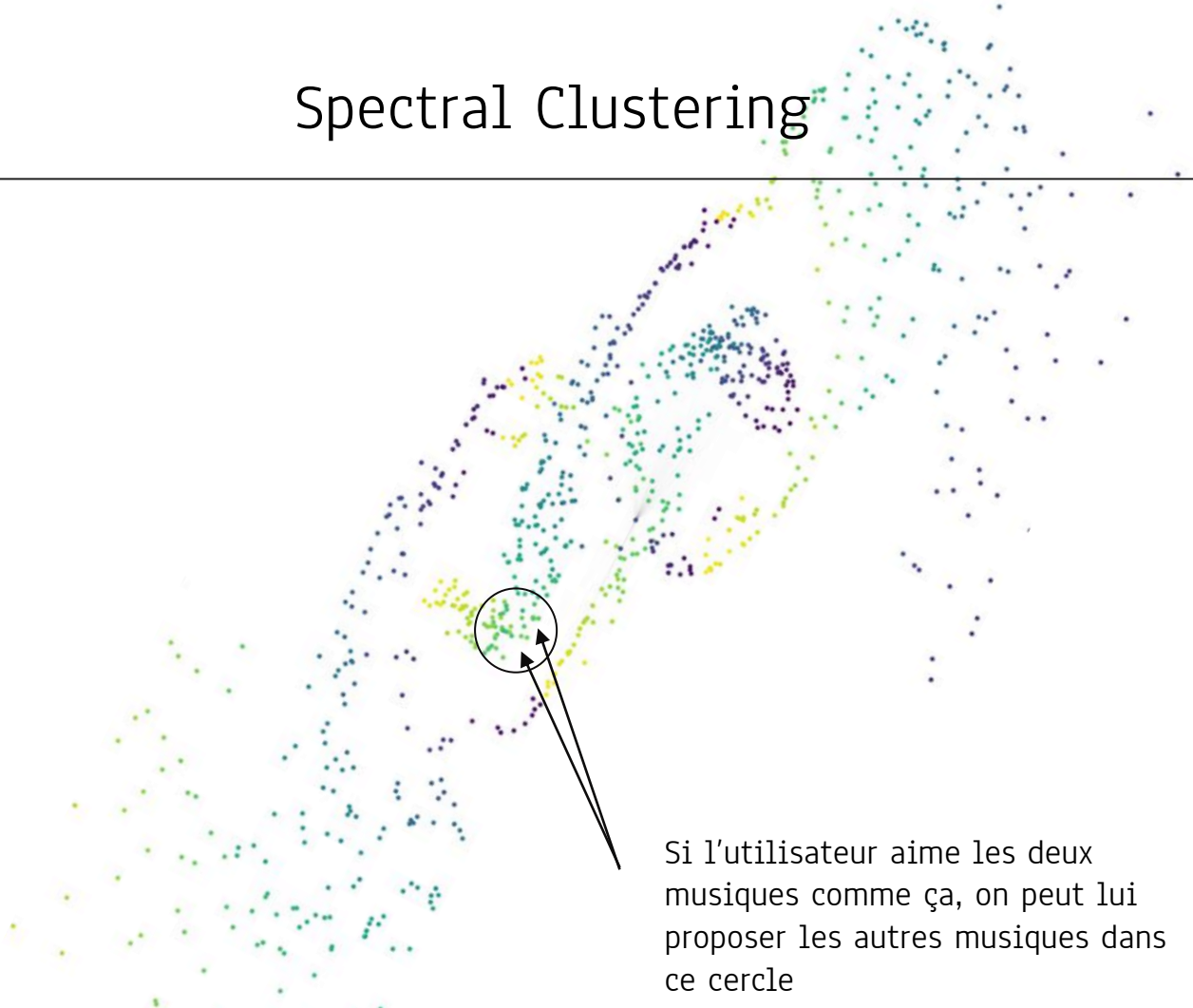


Spectral Clustering:

- Basé sur les librairies networkx et community
- Un point correspond à une musique
- La distance entre deux points correspond au nombre d'utilisateurs qui aiment les deux (juste pour la distance moins de 1)

04

Spectral Clustering

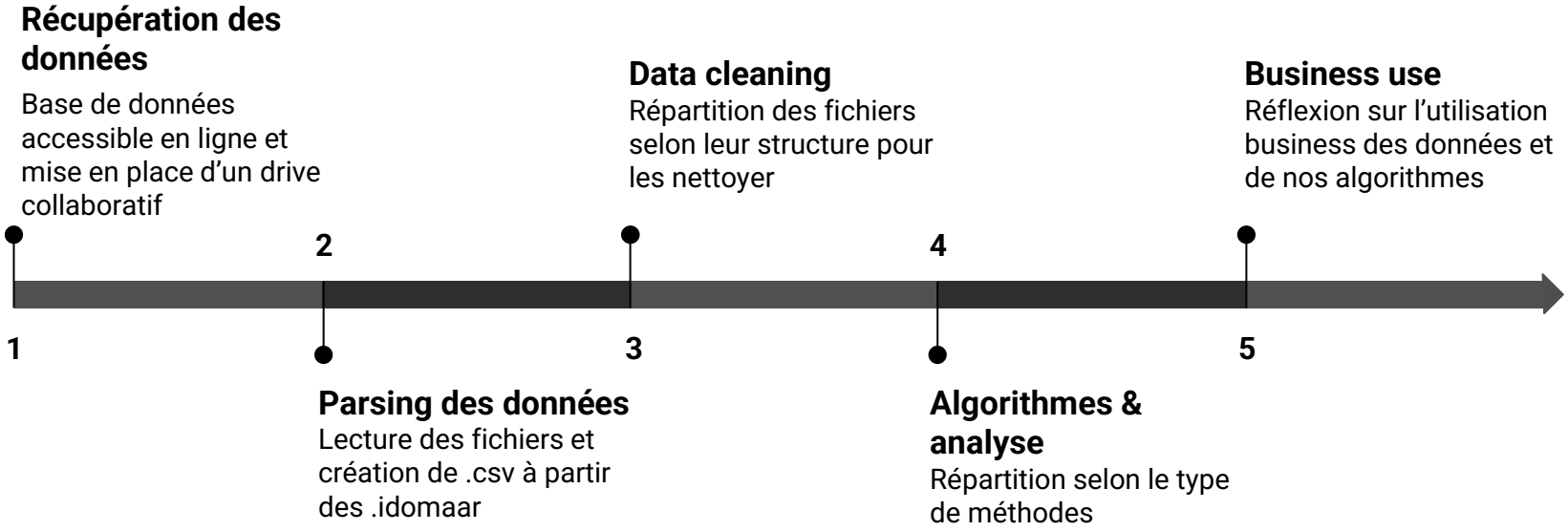


Si l'utilisateur aime les deux musiques comme ça, on peut lui proposer les autres musiques dans ce cercle

SOMMAIRE

05

Organisation



05

Business use (type spotify)

But : satisfaire les envies musicales des utilisateurs pour les fidéliser.

Algorithmes :

- Proposant des musiques écoutées par les personnes du même groupe que l'utilisateur.
- Proposant des musiques d'un autre cluster (mais proche) que celui formé par les méthodes graph-based.
- Création de playlists

Utilité :

- Proposant des musiques à écouter pour un client basé sur ses précédentes écoutes et les musiques qu'il a aimées en fonction des préférences des autres personnes de son cluster.
- Découverte des musiques d'un autre groupe proche du sien.
- Complémentaire avec les utilisations précédentes.

05

Business use (plus original)

Algorithmes :

- Met en relation différentes personnes d'un même cluster

- Déterminer le genre de musique qui plait le plus

Utilité :

- Permet aux utilisateurs de rencontrer des personnes écoutant le même type de musique pour divers buts (concerts...)

- Créer des musiques qui vont plaire à un large groupe (but commercial)

05

Pour aller plus loin …

Avec le même Dataset

Méthode graph based :

- Prise en compte de plus de musiques dans le graph (seulement 500 ici)
- Sélection plus pertinente des musiques (ici celle qui sont le plus aimées)

Autres idées de graph possible :

- Nombre de playlist dans lesquelles les 2 musiques sont présentes
- Nombre de session d'écoute dans lesquelles les 2 musiques sont présentes
- Nombre de tags en commun

Avec un autre Dataset

- Données labellisées pour pouvoir faire de l'apprentissage supervisé
- Données audio pour pouvoir analyser les similitudes

Merci pour votre attention