Europresse, Genius, Youtube et AFS : retour sur quatre expériences de scraping

Corentin Roquebert (ENS, doctorant Centre Max Weber)

Introduction

- Scraping : récupération (plus et surtout moins) automatique de données du web
- Relative « simplicité » technique de ces méthodes dans certains cas
- Travail coopératif
- Usage légal, pirate ou corsaire ?
- Usage raisonné de l'automatisation
- qu'est-ce que l'accessibilité (relative) des données du web fait aux problématiques de recherche ?

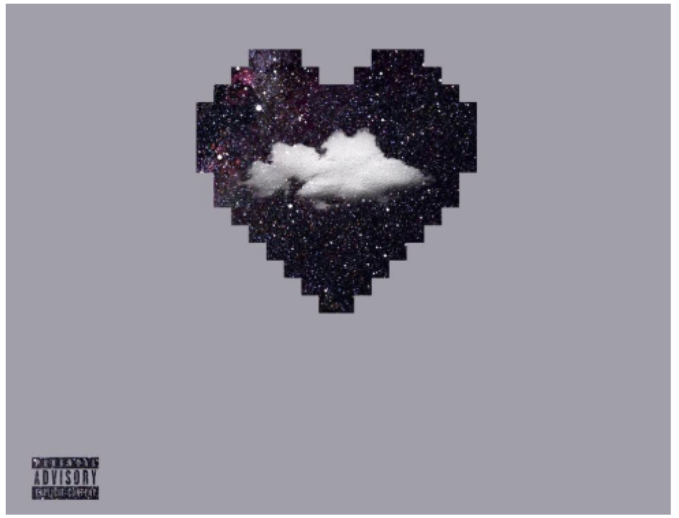
I/ Le scraping comme mise en forme utilisable d'une base de données existantes : l'exemple d'Europresse

- https://quanti.hypotheses.org/1416
- Problème de départ : Europresse donne accès à beaucoup d'articles de presse, mais comment les mettre en forme pour les analyser ?
  - Analyse des métadonnées : date, titre de presse, etc.
  - Analyse textuelle des titres et des textes des articles
- Manuellement, on peut récupérer jusqu'à mille articles répondant à une requête, dans un format HTML
- Script qui permet de transformer le HTML en un format utilisable par des logiciels d'analyse statistique (R, Iramuteq, TXM).
- Il ne s'agit donc pas ici d'un scraping « en ligne » mais de la construction « hors ligne » d'une base de données

II/ Etude du capital social et commercial des rappeurs français (années 2010)

- https://www-cairn-info.acces.bibliotheque-diderot.fr/revue-volume-2020-2-page-61.htm
- Etudier la structuration du monde professionnel du rap français dans les années 2010
- À partir de sources en ligne :
  - Le site Genius, qui recense, entre autres, l'« ensemble » des morceaux de rap français
  - Youtube
- Utilisation de techniques de webscraping pour récupérer des données hétérogènes :
  - Des collaborations entre artistes (featuring) -> 11 400 morceaux, dont 2 600 featurings

- 15/02 : Spider ZED - *Mes ex*
- 14/02 : Krisy - *Paradis d'amour*
- 15/02 : 4keus Gang - *Vois t'as vu*
* 15/02 : Dawa O Mic - *La Galette (Mixtape)*
* 16/02 : Akissa - *Opium - EP*
- 17/02 : Abou Debeing - *Debeinguerie*
- 17/02 : Aladin 135 - *Indigo*
* 17/02 : Le Bon Nob - *P'tit Con*
- 17/02 : Féfé - *Mauve*
* 17/02 : H3RY LÜCK - *FORCE 2*
- 17/02 : Hayce Lemsi - *Électron Libre 2*
* 17/02 : La Massfa - *Frères d'armes*
- 17/02 : Nusky & Vaati - *BLUH*
* 17/02 : Segä - *Nos Futurs*
- 17/02 : Zaho - *Le monde à l'envers*
* 22/02 : JMK$ - *Love & Loyalty*
- 24/02 : A2H - *Les hommes pleurent en hiver (Winter Tape)*
- 24/02 : Jok'Air - *Big Daddy Jok*
- 24/02 : Médine - *Prose Elite*
- 24/02 : O'Trak & ChriStorm - *Sublime abomination*
- 24/02 : Sianna - *Diamant Noir*



Paradis d'amour

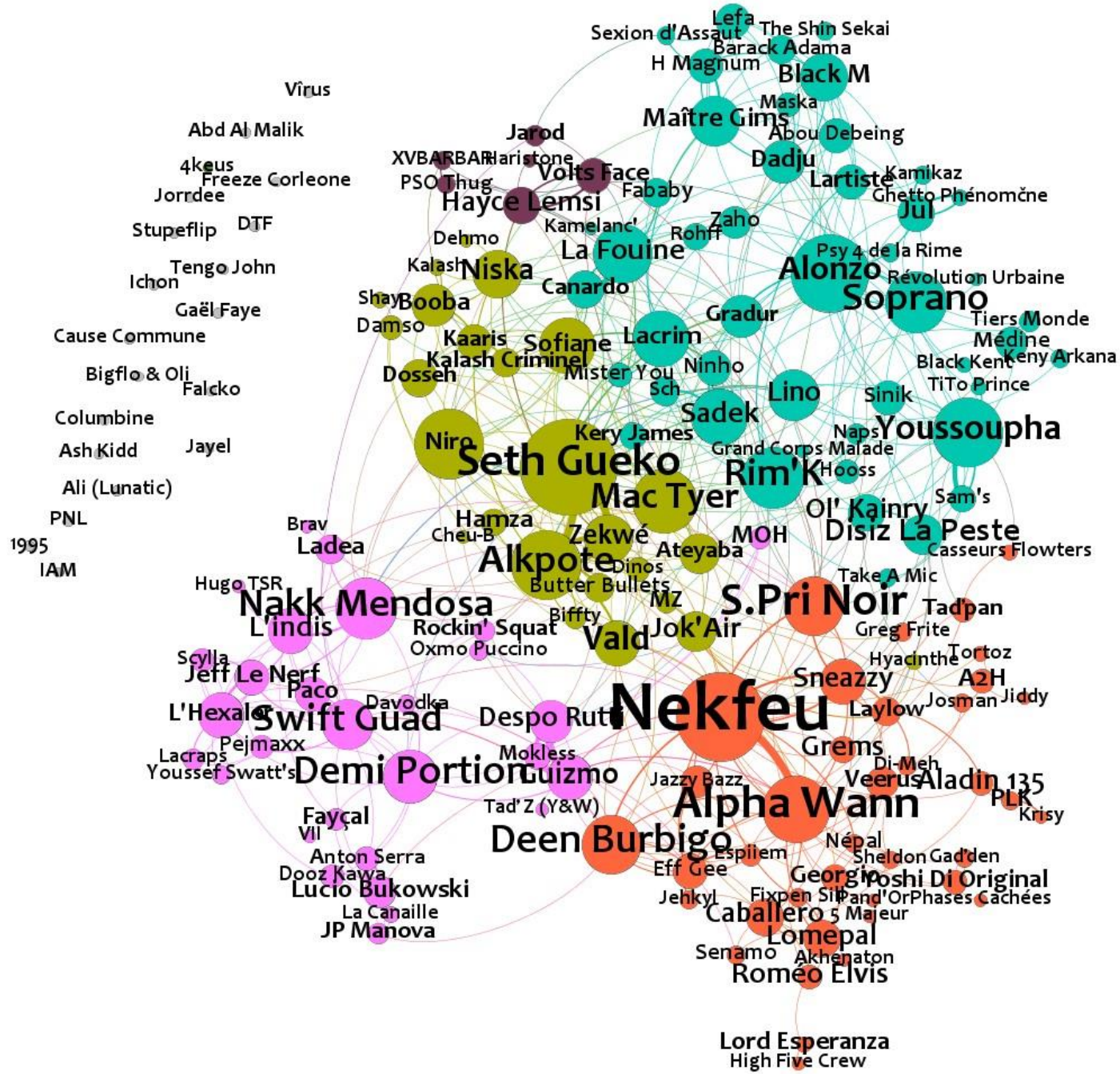
Morceaux :

01 : Vol vers
02 : Discussion nocturne (feat. **Cloe Mailly**)
03 : Blessed
04 : Paradis d'amour
05 : Erotiquement votre [Interlude] (feat. Instagram followers)
06 : Ils pensent
07 : Elle, me parle
08 : La vérité sur terre (feat. **Koola**)

Genius

- Scraping « brut » plutôt qu'usage de leur API
- Représentation imparfaite de l'activité du rap français :
  - Biais de popularité : surreprésentation des artistes les plus renommés
  - Biais de genre : sous-représentation des rappeuses
  - Biais temporel : surreprésentation des productions récentes
- Des décisions dans les recodages
- Permet de
  - Distinguer différents pôles de collaboration dans le rap français
  - D'identifier différentes logiques d'usage du capital social des rappeurs : connexité, intégration, externalisation, densité.
- Faire les liens avec les textes : associer chaque couplet à un artiste

Taille du nœud : nombre d'invitations données
Taille du texte : nombre d'invitations reçus



La construction d'un indicateur de popularité sur Youtube

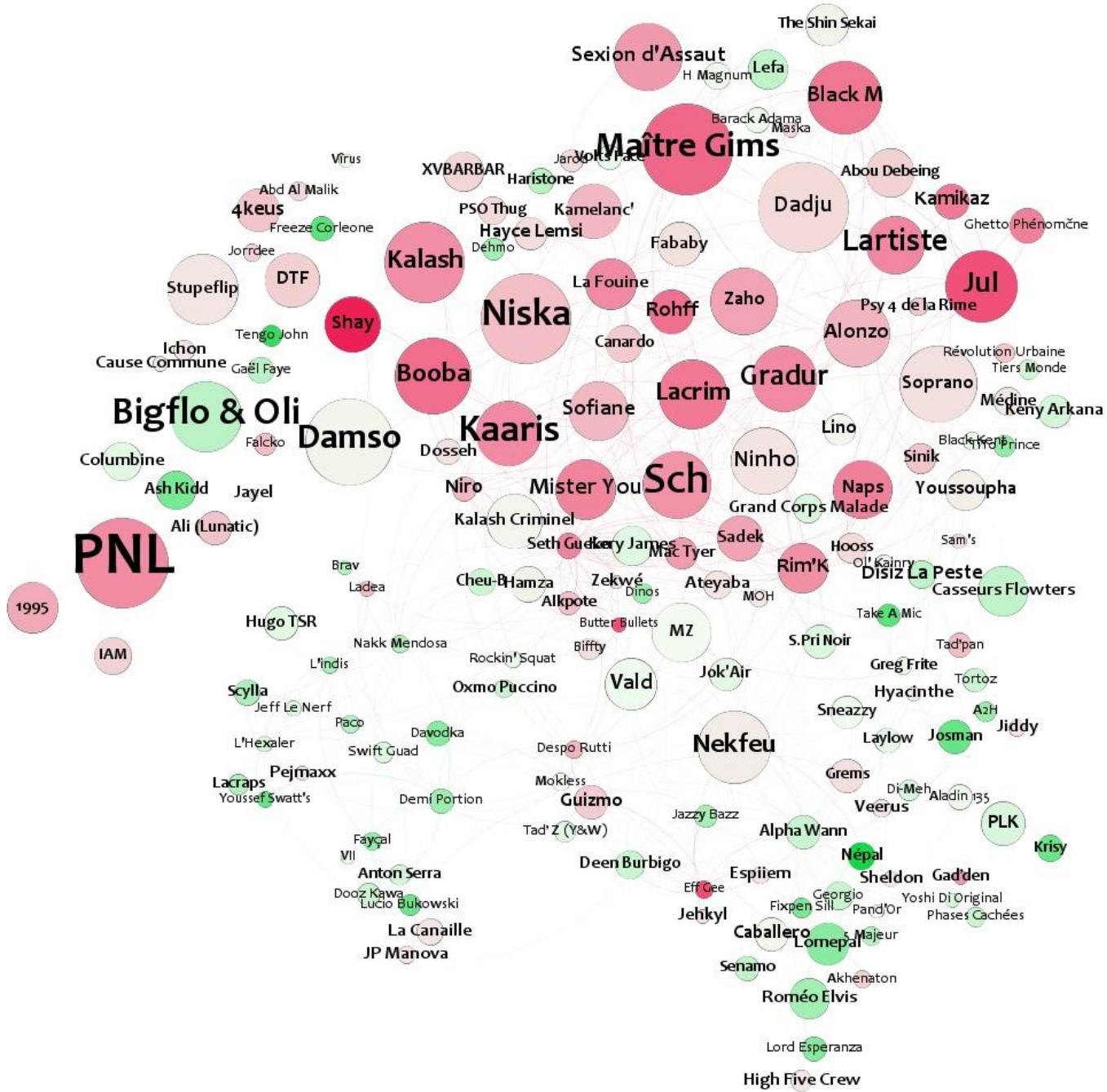
- But : récupérer des vidéos contenant des morceaux de chaque artiste
- Scraping de Youtube :
  - Sur les 200 rappeurs avec le plus de morceaux recensés sur Genius
  - Requête : « nom du rappeur + rap »
  - 400 premiers résultats de chaque requête
  - Pour éviter d'être « striké », long processus : script durant deux jours pour récupérer en théorie 80 000 résultats
  - Au final 10 000 vidéos utilisables
- Pour chaque vidéo, on récupère :
  - Titre
  - Date
  - Durée

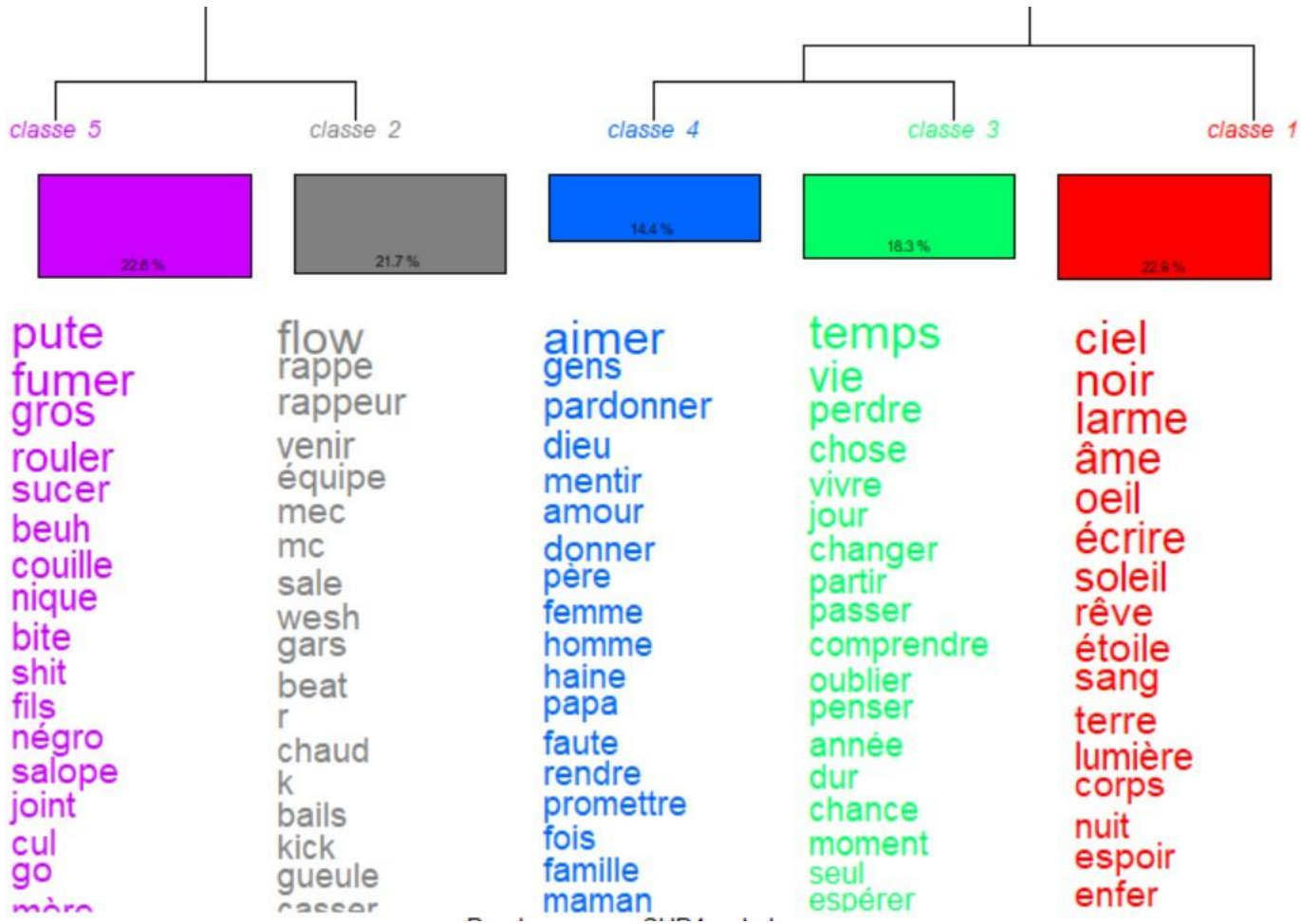
Un long processus de recodage manuel

- De façon automatique, on enlève :
  - Les doublons
  - Les vidéos avec peu de vues
  - Les vidéos trop courtes ou trop longues
  - De 27 000 à 13 500 vidéos
- A la main :
  - Semi-automatisé : enlever les instrumentales, les interviews, les reprises etc.
  - Mais surtout à la main : vérification qu'il s'agit bien de vidéos musicales avec le rappeur en question
  - 10 000 vidéos

Recodage et construction d’indicateurs

- A partir des titres bruts des vidéos :
  - Clip
  - Freestyle
  - Featuring
- A partir des chaînes :
  - Chaîne de radio
  - Chaîne officielle de chaque artiste
- Indicateur plus précis du featuring, quatre modalités :
  - Morceau solo (un seul artiste de la base dans le titre, pas de marque de featuring)
  - Featuring interne (plusieurs artistes de la base dans le titre + marque de feat)
  - Featuring inconnu (un seul artiste de la base + marque de feat)

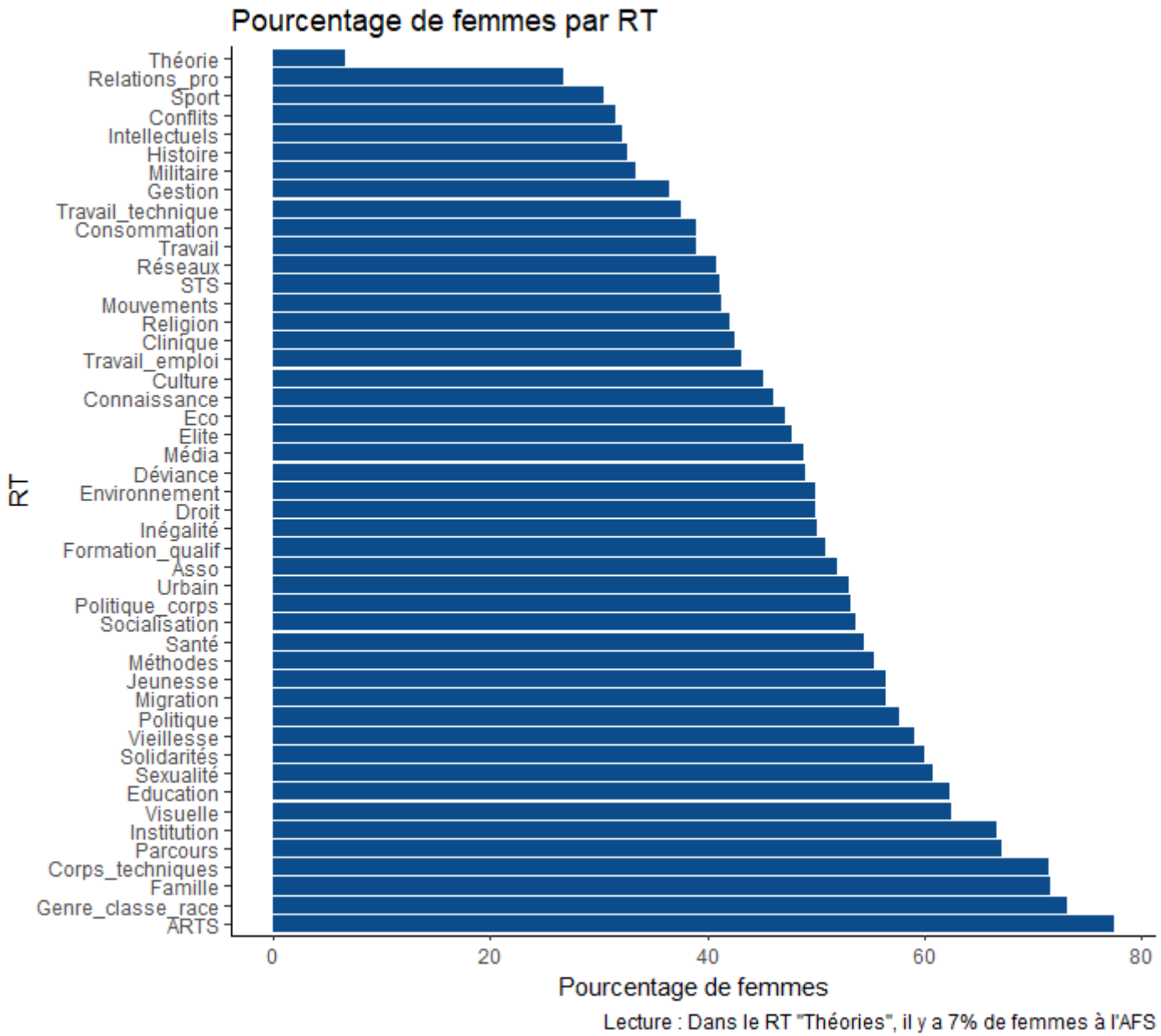

The Shin Sekai
Sexion d'Assaut
H Magnum
Lefa
Black M
Barack Adama
Maska
Maître Gims
Vîrus
Jardin Volts Face
XVBARBAR
Haristone
Abou Debeing
Kamikaz
Abd Al Malik
4keus
Freeze Corleone
PSO Thug
Kamelanc'
Hayce Lemsi
Dadju
Ghetto Phénomčne
Lartiste
Jorrdee
Dehmo
Fababy
Kalash
Stupeflip
DTF
La Fouine
Zaho
Jul
Niska
Rohff
Psy 4 de la Rime
Shay
Tengo John
Canardo
Alonzo
Ichon
Cause Commune
Gaël Faye
Booba
Révolution Urbaine
Tiers Monde
Soprano
Lacrim
Gradur
Bigflo & Oli
Sofiane
Médine
Keny Arkana
Damso
Kaaris
Lino
Falcko
Dosseh
Ninho
Sinik
Columbine
Ash Kidd
Jayel
Niro
Mister You
Sch
Naps
Youssoupha
Grand Corps Malade
Kalash Criminel
Ali (Lunatic)
PNL
Seth Gueko
Kery James
Mac Tyer
Sadek
Hooss
Sam's
Rim'K
Ol' Kainry
Disiz La Peste
Brav
Ladea
Cheu-B
Hamza
Zekwé
Dinos
Ateyaba
MOH
Casseurs Flowters
Alkpote
1995
Take A Mic
Hugo TSR
Butter Bullets
MZ
S.Pri Noir
Nakk Mendosa
Biffty
Tad'pan
IAM
L'indis
Rockin' Squat
Greg Frite
Jok'Air
Vald
Tortoz
Oxmo Puccino
Hyacinthe
Scylla
Jeff Le Nerf
Sneazzy
A2H
Jiddy
Paco
Davodka
Laylow
Josman
L'Hexaler
Swift Guad
Despo Rutti
Nekfeu
Grems
Pejmaxx
Mokless
Lacraps
Youssef Swatt's
Demi Portion
Guizmo
Di-Meh
Aladin
135
Veerus
Jazzy Bazz
PLK
Tad' Z (Y&W)
Alpha Wann
Krisy
Fayçal
VII
Népal
Deen Burbigo
Espiiem
Sheldon
Gad'den
Anton Serra
Eff Gee
Georgio
Yoshi Di Original
Dooz Kawa
Lucio Bukowski
Fixpen Sill
Pand'Or
Phases Cachées
Jehkyl
La Canaille
Caballero
JP Manova
Majeur
Lomepal
Senamo
Akhenaton
Roméo Elvis
Lord Esperanza
High Five Crew


classe 5
classe 2
classe 4
classe 3
classe 1
22.6 %
21.7 %
14.4 %
18.3 %
22.9 %
pute
fumer
gros
rouler
sucer
beuh
couille
nique
bite
shit
fils
négro
salope
joint
cul
go
flow
rappe
rappeur
venir
équipe
mec
mc
sale
wesh
gars
beat
r
chaud
k
bails
kick
gueule
aimer
gens
pardonner
dieu
mentir
amour
donner
père
femme
homme
haine
papa
faute
rendre
promettre
fois
famille
maman
temps
vie
perdre
chose
vivre
jour
changer
partir
passer
comprendre
oublier
penser
année
dur
chance
moment
seul
espérer
ciel
noir
larme
âme
oeil
écrire
soleil
rêve
étoile
sang
terre
lumière
corps
nuit
espoir
enfer

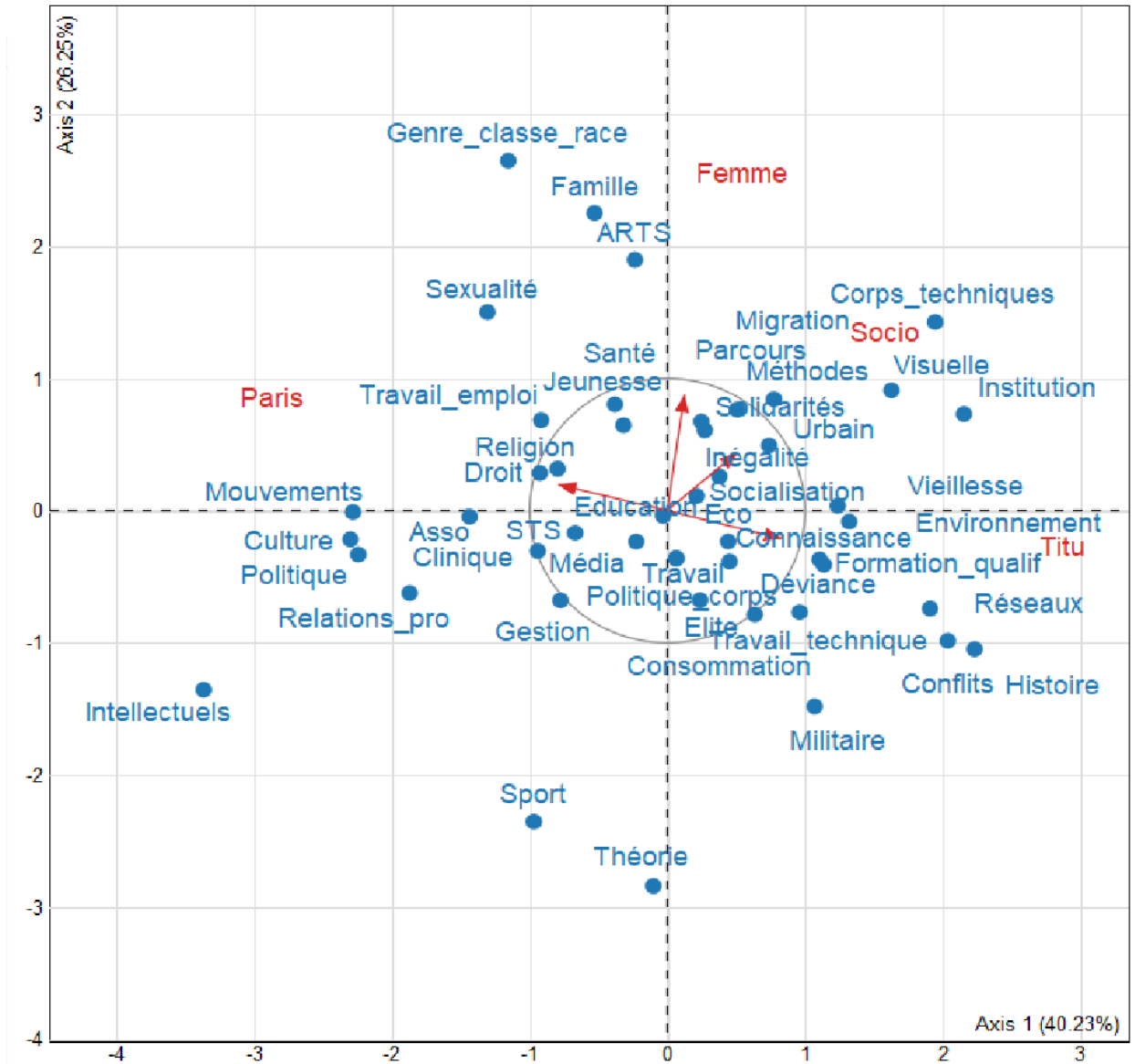
- Dans le cas présenté ici, l'utilisation de données Youtube permet trois choses :
  - Confirmer ou mettre en doute des intuitions sur le rôle des featuring dans l'économie général du monde professionnel du rap : les featurings entre professionnels du rap sont rémunérateurs commercialement
  - Faire apparaître une opposition dans les modes de collaboration entre des types différents de rappeur
  - Poser de nouvelles questions sur les liens entre formes de popularité, formes esthétiques et collaboration artistique.

III/ Scraping et constitution de bases sur le congrès de l'Association française de sociologie (2019)

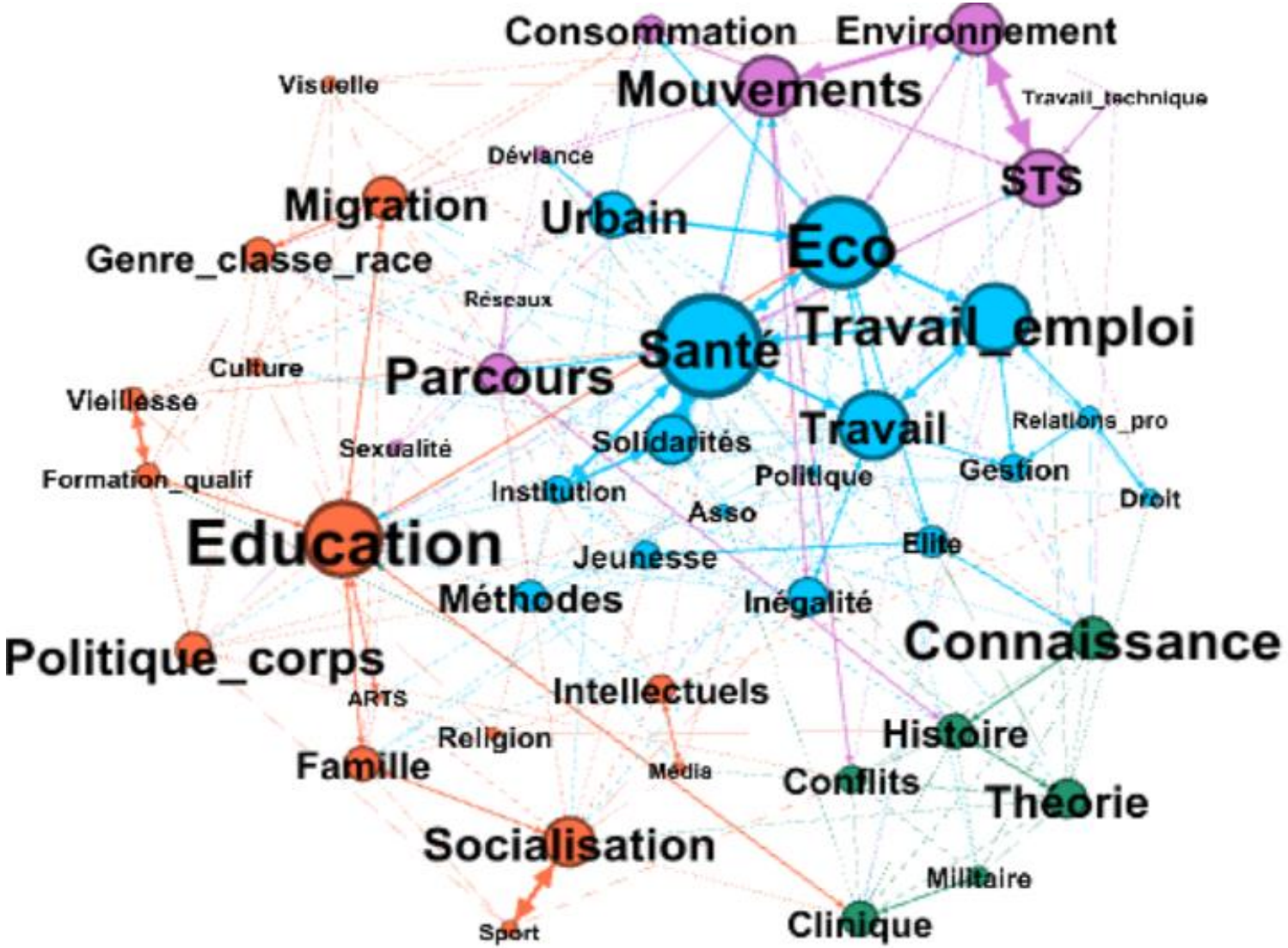
- Constitution d'une base de données :
  - Scraping du programme de l'AFS
  - Récolte manuelle d'informations en ligne sur les 1284 participant.e.s du congrès.
- Refus (compréhensible) de la part de l'AFS de me donner accès :
  - A une version directement utilisable de leur programme
  - A des informations non visibles sur le programme
- Scraping permet ici de mettre en forme rapidement une information déjà existante (le programme de l'AFS)
- Une base des communications, avec les informations suivantes : auteurs, RT, titre, texte.
- Une base des participants


Pourcentage de femmes par RT
RT
Théorie
Relations_pro
Sport
Conflits
Intellectuels
Histoire
Militaire
Gestion
Travail_technique
Consommation
Travail
Réseaux
STS
Mouvements
Religion
Clinique
Travail_emploi
Culture
Connaissance
Eco
Elite
Média
Déviance
Environnement
Droit
Inégalité
Formation_qualif
Asso
Urbain
Politique_corps
Socialisation
Santé
Méthodes
Jeunesse
Migration
Politique
Vieillesse
Solidarités
Sexualité
Education
Visuelle
Institution
Parcours
Corps_techniques
Famille
Genre_classe_race
ARTS
0
20
40
60
80
Pourcentage de femmes
Lecture : Dans le RT "Théories", il y a 7% de femmes à l'AFS

Analyse en composante principale (ACP) à partir des variables de composition des RT



Le réseau des RT à partir des multicommunications



Lien : participant.e.s qui communiquent dans les deux RT (sans les sessions croisées)
Taille des nœuds : degré pondéré
Taille des textes : centralité d'intermédiarité
Couleur : cluster (algorithme de Louvain)

Conclusion : Différents usages du scraping

- Scraping pour résoudre un problème technique de mise en forme des données (Europresse)
- Scraping pour accéder à une information :
  - Qu'on n'aurait pas autrement (RapGenius)
  - Qui serait trop longue à formaliser à la main (programme du congrès de l'AFS)
- Accès à des méthodes qui vont :
  - Répondre de manière originale à des problématiques de recherche
  - Mais aussi amener des nouvelles questions de recherche