Analyse des tweets

Data science : Clustering

Encadrant : Lazhar Labiod

Agliz Yasmine / Haddad Ayale / Plançon Alexandra

2019-2020

Table des matières

INTRODUCTION

Ce projet porte sur l'extraction et analyse des tweets. Il a été proposé par Monsieur Lazhar Labiod dans le cadre de l'UE de Data science 2.

Ce groupe se compose de trois étudiants : Agliz Yasmine, Haddad Ayale et Plançon Alexandra. Le but de ce projet est d'analyser des tweets et d'en extraire des informations à partir d'outils de clustering étudiés en cours

Tout d'abord, nous allons commencer par expliquer le processus d'extraction des tweets grâce à l'API de Tweeter.

Ensuite, nous allons détailler la partie de traitements des tweets qui nous permets d'obtenir une matrice de contingence pour la suite de l'analyse. Puis nous effectuons une analyse descriptive des datas obtenues.

Pour finir, nous expliquerons les méthodes d'analyse factorielle et de clustering employées ainsi que les résultats obtenus.

EXTRACTION DES TWEETS

Pour cette première étape nous avons utilisé les API publiques de Twitter. L'un des avantages d'utiliser ces APIs est la fiabilité des données collectées. Un autre avantage de cette interface publique, est que nous n'avons pas été obligés de fournir nos clés de compte Twitter à un quelconque service tierce. Toutefois, en utilisant un accès public aux APIs, nous avons été confrontés aux restrictions imposées par Tweeter. Par exemple, nous ne pouvons pas collecter plus 450 tweets par recherche. De plus, il faut attendre 15 min entre chaque appel de l'API.

Les tweets ont été collectés à l'aide de la librairie tweepy. Ce package permet de récupérer les tweets en précisant le terme qu'un tweet doit contenir. Pour ce faire nous avons eu besoin de clés d'authentification obtenues via Twitter.

La collecte a été faite en 5 étapes. Nous avons récupéré 900 tweets de 5 thèmes de musique différents : Rock, Jazz, Rap, Pop, Reggae. Nous avons ainsi collecté 4500 tweets que nous avons enregistré dans le fichier « all_type_music.json ».

Beaucoup des informations extraites ne nous était pas utiles ou du moins inexploitables.

Prenons comme exemple la localisation, de nombreux utilisateurs ont mal renseigné leur localisation voire pas du tout. Dans l'exemple ci-dessous nous pouvons remarquer que la localisation n'a pas de format particulier ce qui complique l'extraction des informations de cette variable.

user_location
Los Angeles, CA
sna frnaisco
Jamaica, NY
LA

LES LOCALISATION DES 5 PREMIERS TWEETS

Les seules variables retenues est le contenu syntaxique des tweets.

PRETRAITEMENT

Le prétraitement que nous effectuons dans ce projet consiste à préparer la base de données collectée pour en extraire une matrice de contingence exploitable par le reste de notre analyse. Dans un premier temps, nous procédons à un nettoyage au niveau de la base de données des tweets :

1. Suppression des tweets ayant le texte tronqué
2. Suppression des données en doublon dû à un retweet par exemple
3. Suppression des tweets n'ayant que des mentions d'utilisateurs ou url

Dans un deuxième temps, nous appliquons quelques traitements afin d'obtenir un vocabulaire de mots sans redondance. Pour ce faire, on utilise un Tockenizer afin d'extraire la liste de termes employée dans les tweets. Puis, nous procédons à deux traitements principaux :

La **lemmatisation** qui désigne un traitement lexical qui consiste à transformer les verbes, adjectifs et substantifs en leur forme canonique que l'on désigne sous le terme de lemme.

La **racinisation** ou **désuffixation** qui est un procédé de transformation des flexions en leur radical ou racine.

D'autres opérations effectuées lors du nettoyage du vocabulaire :

1. Suppression des mentions
2. Supression des émoticones
3. Conversion des textes en minuscule
4. Supression des URLS
5. Supression des noms d'utilisateurs cités
6. Supression des # dans les hashtags
7. Supression des caractères répétés (bonjourrrr devient bonjour)

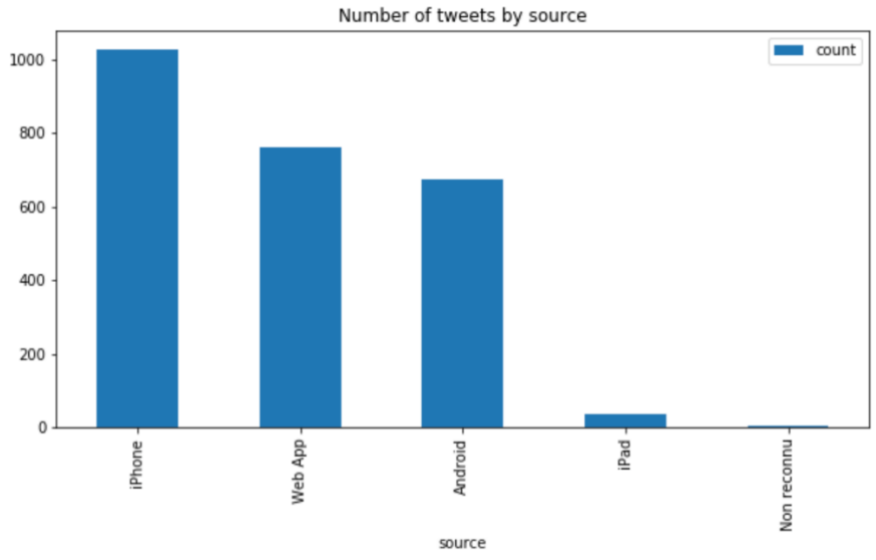
ANALYSE DESCRIPTIVE DES DONNEES

EXPLORATION DESCRIPTIVE

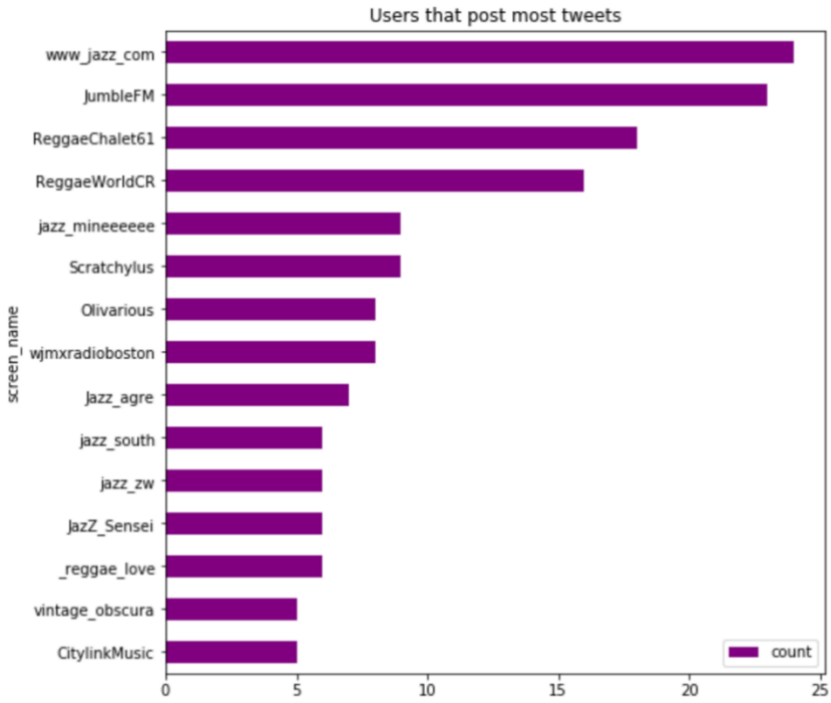
Le nombre de tweets non tronqués est de 3676, après le nettoyage il ne reste plus que 2507 tweets. Dans l'image ci-dessous on peut visualiser d'autres données intéressantes.

```
Total tweets no trunced: 3676

Le nombre unique d'auteurs est de 3230
Le nombre de tweets originaux est de 1463
Le nombre de doublons de tweets est de 1176

Mean retweets: 0.6

Top 5 RTed tweets:
------------------
RT @AnaCookman: Rock 'n' Roll pioneer Chuck Berry was born in 1926. Berry's music was a major influence on The Beatle
s, AC/DC and the Rolli… - True
RT @wavycoma: when yo face pop up in my head, all i can say is damn.. - True
RT @hefferk: Day 106 of New project called 'Where I live' Dun Laoghaire Harbour @Photooftheday @dublin @PhotosOfDubli
n @VisitDublin @OldeEi… - True
RT @YungGotThejuice: when i hear Pop Smoke cough in @liltjay Zoo York😔😔💫🕊️ https://t.co/HIwP8ezC2j - True
RT @wesstreeting: If you got the wrong end of the stick, or just chose to have a cheap pop, maybe share this... - Tru
e
```

De plus, on a pu constater que la plupart des tweets sont émis à partir d'un smartphone, avec un taux plus élevé pour l'iPhone.



Nous avons remarqué que dans les tweets collectés, certains noms d'utilisateurs reviennent souvent. Le graphique ci-dessous représente les utilisateurs qui ont posté le plus de tweets. Ce classement permet de donner un premier point de vue sur le classement des auteurs.



Un hashtag, est représenté par le symbole #, il sert à indexer des mots-clés ou des sujets sur Twitter. Cette fonctionnalité a été créée pour permettre aux utilisateurs de suivre facilement des sujets qui les intéressent. Ainsi on a récupéré les top 10 des hashtag des tweets pour observer les tendances des sujets choisis. Sans surprise, on peut observer les différents thèmes de musique (jazz, reggae, rap). On remarque notamment le terme radio dans le top 10. Nous pouvons en conclure que les utilisateurs ont tendance à parler de musique qu'ils écoutent à la radio.

```
Top 10 hashtags:
----------------
nowplaying - 111
reggae - 42
jazz - 29
listen - 25
music - 24
radio - 22
1 - 17
howmanytimes - 14
rap - 12
np - 10
```

Une autre information qui pourrait être intéressante est le top 10 des mentions d'utilisateurs dans les tweets. On peut remarquer que le compte de YouTube est le plus souvent mentionné.

```
Top 10 mentions:
----------------
youtube - 41
wwwjazzcom - 23
originalfunko - 13
addiselfgh - 12
billboard - 12
promodj - 11
reggaeworldcr - 9
abhisarsharma - 6
artistrack - 6
thejazzsoul - 6
```

ANALYSE DE SENTIMENT

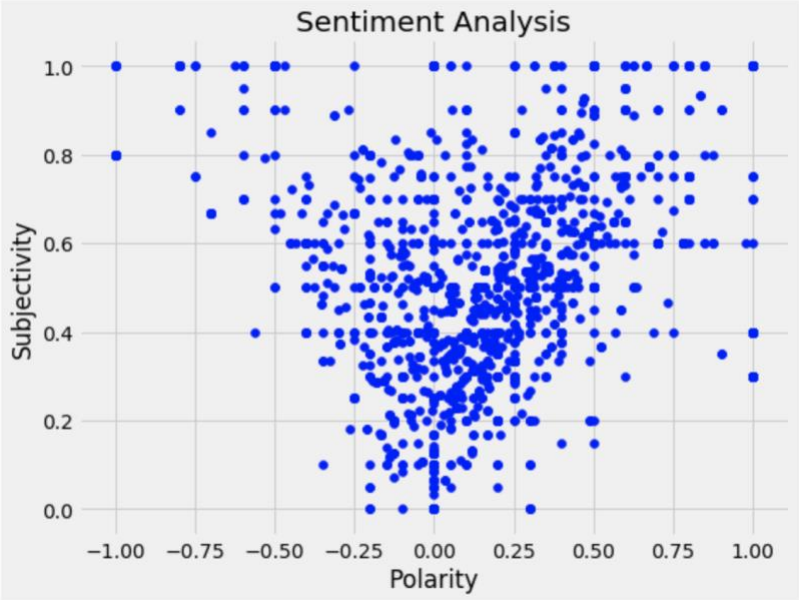
L'analyse des sentiments, également appelée « Opinion Mining », fait référence à l'utilisation du traitement du langage naturel pour déterminer l'attitude, les opinions et les émotions d'un locuteur. C'est le processus qui permet de déterminer si un écrit est positif ou négatif. Nous appelons ce procédé la polarité du contenu. En tant qu'êtres humains, nous sommes capables de classer le texte en positif ou en négatif inconsciemment.

Prenons comme exemple, la phrase « L'enfant avait un magnifique sourire sur son visage ». Celle-ci nous donnera très probablement un sentiment positif. Une personne aboutis à une telle conclusion en examinant les mots et en faisant la moyenne des termes portants un sens positifs et des termes portants négatifs. Par exemple, les mots « magnifique » et « sourire » sont considérés comme étant positifs, tandis que des mots comme « le », « enfant » et « visage » sont neutres. Par conséquent, le sentiment général de la phrase est considéré comme étant positif.

Une utilisation courante de cette technologie provient de son déploiement dans l'espace des réseaux sociaux. En effet, cela permet d'acquérir l'opinion du public sur certains sujets.

Nous avons utilisé la librairie TextBlob pour l'analyse de sentiment. On peut observer que la majorité des tweets collectés porte plus des sentiments positifs que négatifs.

De plus, nous avons également pu estimer la subjectivité des sentiments. Dans le graphe ci-dessous, on peut remarquer que la plupart des sentiments exprimés ne sont pas subjectifs.



Nous retrouvons 41% des tweets collectés qui sont positifs et seulement 13% négatifs. La plupart des tweets semblent être neutres. Cela pourrait être expliqué par les tweets que le package TextBlob n'arrive pas à traiter de manière efficace, à cause des acronymes de mots par exemple.

EXPLORATION MULTIDIMENSIONNELLE

Nous souhaitons dans cette section appliquer les méthodes de data mining aux tweets. Il faut pour cela que nous les transformions en matrice de données numériques.

La représentation en sac de mots (bag-of-words) est la plus utilisée pour ce genre de problématique. Dans un premier temps, nous avons construit une matrice de contingence tweet-terme. Chaque case de cette matrice contient l'occurrence du terme dans le tweet. Au vu des résultats obtenus lors du clustering avec cette matrice, nous avons décidé d'utiliser un autre score que la fréquence qui est le score TF-IDF. Celui s'est avéré être plus efficace pour la représentation tweets-termes.

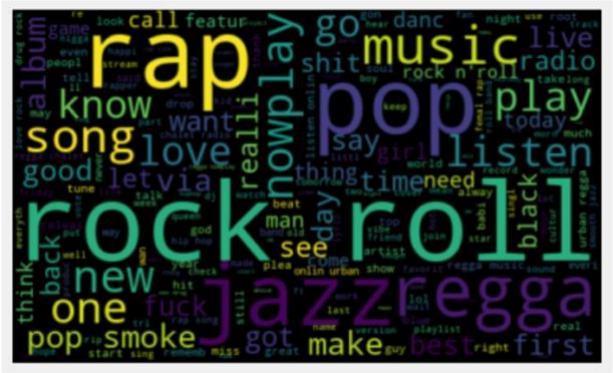
Le score TF-IDF est une méthode de pondération. Cette mesure statistique permet d'évaluer l'importance d'un terme contenu dans un document, relativement au corpus. Le poids d'un terme augmente proportionnellement au nombre d'occurrences du mot dans le document. Il varie également en fonction de la fréquence du mot dans le corpus.

Pour des raisons de visibilité nous ne pouvons malheureusement pas travailler sur 2000 tweets. Nous avons donc décidé de classer les termes en fonction de leur score TF-IDF. Puis nous avons sélectionné les 450 top termes. On suppose qu'un terme est pertinent si la somme des valeurs de TF-IDF pour tous les tweets réunis sont élevés.

id	rock	roll	rap	regga	pop	jazz	music	song	like	nowplay
0	0.098551	0.101521	0.0	0.0	0.0	0.000000	0.143324	0.0	0.0	0.000000
1	0.140800	0.145044	0.0	0.0	0.0	0.000000	0.204768	0.0	0.0	0.225246
2	0.163063	0.167978	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000
3	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000
10	0.143222	0.147538	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.000000

ÉCHANTILLON DE LA MATRICE DOCUMENT-TERME

Pour une meilleure visualisation on a utilisé wordcloud qui est une représentation visuelle des termes dans les tweets les plus utilisées. Bien entendu, il est totalement logique que les 5 styles musicaux apparaissent dans les termes les plus fréquents. Ces termes nous serviront de repère dans la suite.



ANALYSE FACTORIELLE DES CORRESPONDANCES

L'analyse factorielle des correspondances est une extension de l'analyse en composantes principales pour analyser l'association entre deux variables qualitatives. Dans notre cas ces deux variables sont : les tweets et les termes qui les composent.

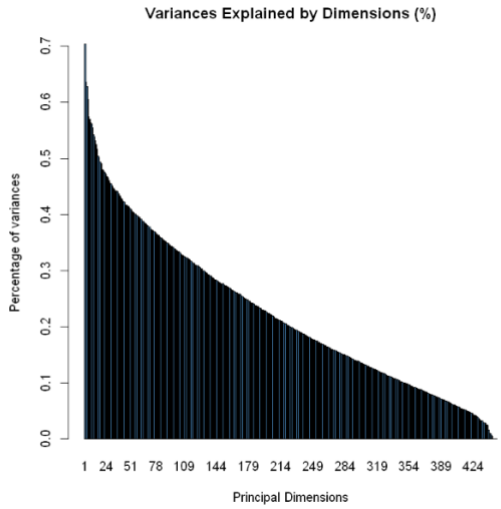
L'AFC nous permet de résumer et de visualiser l'information contenue dans le TABLEAU DE CONTINGENCE formé par les deux variables catégorielles. Le tableau de contingence contient les scores TF-IDF des termes dans les tweets.

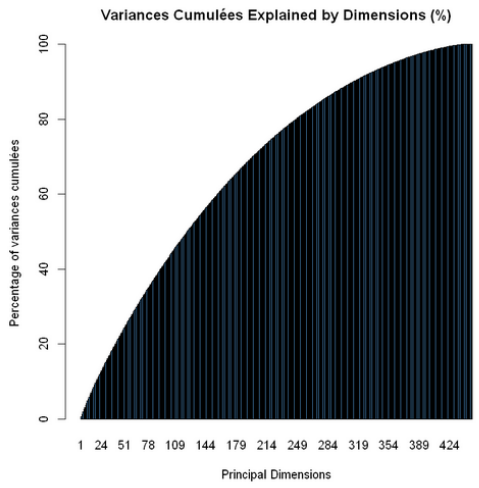
L'AFC retourne les coordonnées des éléments des colonnes et des lignes du tableau de contingence. Ces coordonnées permettent de visualiser graphiquement l'association entre les termes et les tweets.

Le résultat attendu est d'avoir des tweets portants sur le même style musical dans la même direction que les mots du jargon de ce même style.

RESULTATS

Nous avons donc effectué une AFC sur la matrice de contingence. Nous avons obtenu les résultats suivants :

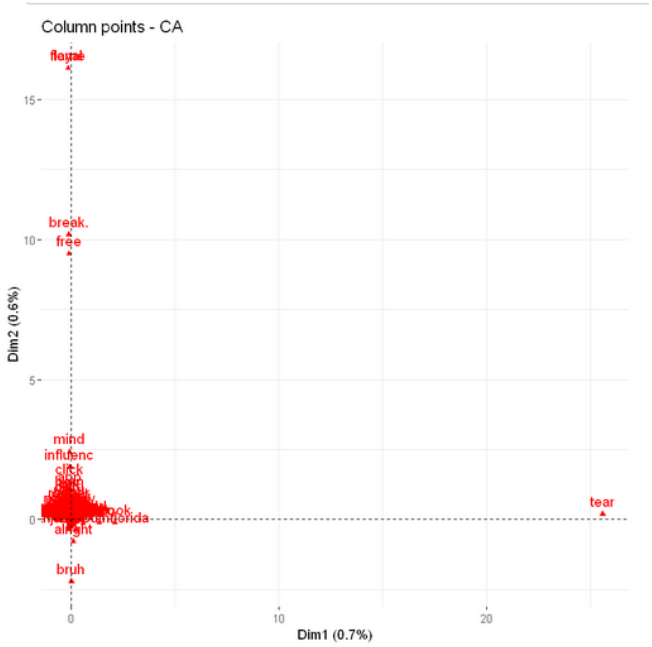




Nous pouvons remarqué que la variance cumulées expliquée croit de manière logarithmique donc pas assez rapidement. En effet, la variance expliquée atteint 80% qu'à la dimension 249 cela réduit de 50% le nombre de dimension.

Malheureusement, la représentation des variables sur deux dimensions reste ininterprétables à cause du grand nombre de variables et des observations. En effet, le nombre de dimensions utilisé n'est pas assez grand pour représenter les datas. Deux dimensions ne representent que 1,34% de l'information globale des observations.

REPRESENTATIONS DES VARIABLES :



Les données ne sont toujours pas interprétables visuellement. Cela dit grâce aux coordonnées des variables on peut vérifier la cohérence des résultats obtenus.

La dimension 1 et 2 ont du mal à bien séparer les termes entre eux. Prenons donc comme exemple les cordonnées de quelques variables sur la dimension 3 :

Rock: -0.0545576372216395

Roll: -0.0566541455232552

Beatl (Beattles): -0.0559684586105207

Le **rock 'n'roll** est un genre musical, ayant émergé aux États-Unis à la fin des années 1940. **The Beattles** est l'un des groupes les plus connus du rock'n'roll.

Ac: -0.0699215436733911

Dc : -0.0699215436733912

AC/DC est un groupe de hard rock australien.

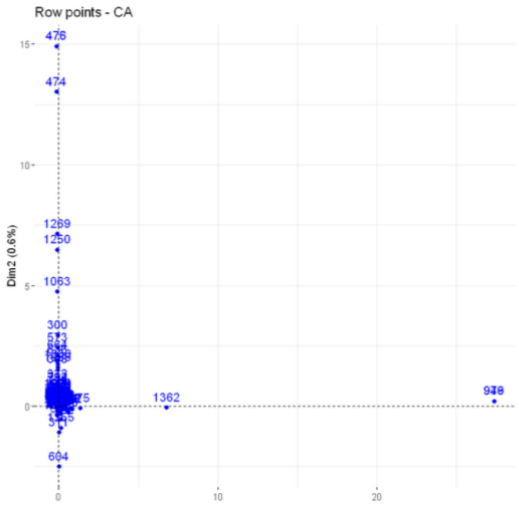
Bob : -0.1139680759212

Marley : -0.110294351197977

Bob Marley, est le chanteur le plus connu du reggae

Nous pouvons remarquer que la dimension 3 rassemble bien les mots ayant le même contexte musical.

REPRESENTATIONS DES OBSERVATIONS



En ce qui concerne les tweets, nous pouvons remarquer que la dimension 2 arrive bien à représenter quelques tweets notamment les tweets 1250,1269 et 1063 :

1250 -> “If you truly love reggae contribution”

1269 -> "@Martindj_marto- random reggae 6 (quarantine edition 🔥🔥🔥🔥🔥🔥🔥 #mohspice”

1063 -> “RT @efyastacey: I’m just imagining how the tune is gonna be?? Is going to be Azonto, hip hop, Rnb, high life, hip life, reggae, Raggae, blu…”

Ces trois tweets portent tous sur le style musical du reggae.

REPRESENTATION GLOBALE :

Nous ne pouvons malheureusement pas vérifier la corrélation entre les variables et les tweets avec le cercle de corrélation car celui-ci est illisible.

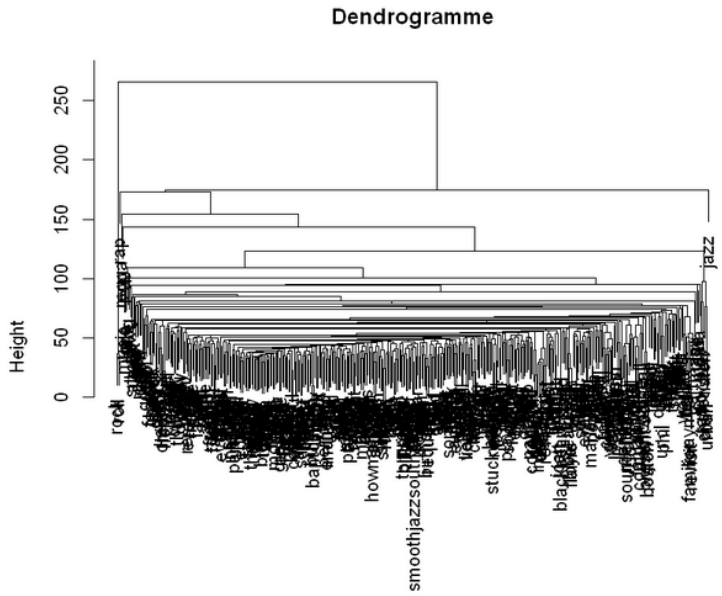
CLASSIFICATION ASCENDANTE HIERARCHIQUE

La classification ascendante hiérarchique (CAH*)* est une technique statistique visant à partitionner une population en différentes classes ou sous-groupes. Dans notre cas nous allons essayer de partitionner les termes en sous-groupes.

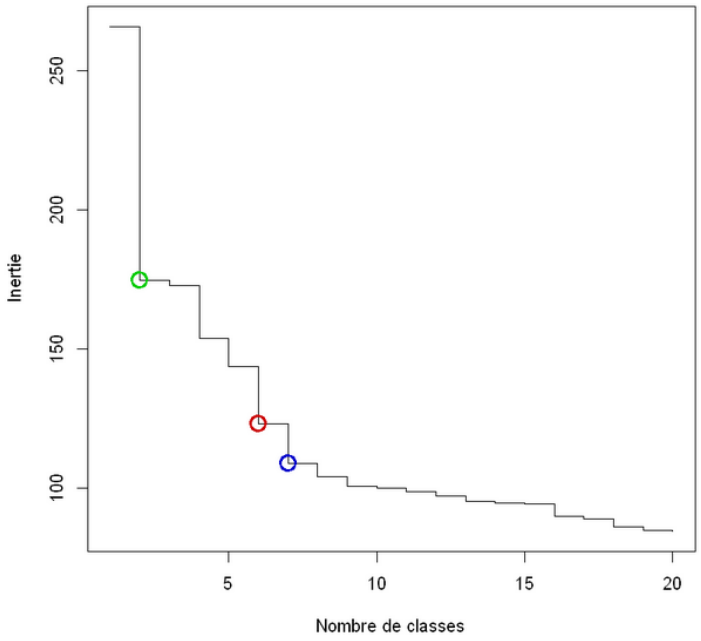
On cherche à ce que les termes regroupés au sein d'une même classe soient le plus semblables possibles tandis que les classes soient le plus dissemblables. Autrement, on cherche à maximiser l'homogénéité intra-classe et maximiser l'hétérogénéité inter-classe.

Le principe de la CAH est de rassembler des termes selon un critère de ressemblance défini au préalable qui s'exprimera sous la forme d'une matrice de distances, exprimant la distance existante entre chaque terme pris deux à deux. Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante. La CAH va ensuite rassembler les individus de manière itérative afin de produire un dendrogramme ou arbre de classification.

DENDROGRAMME

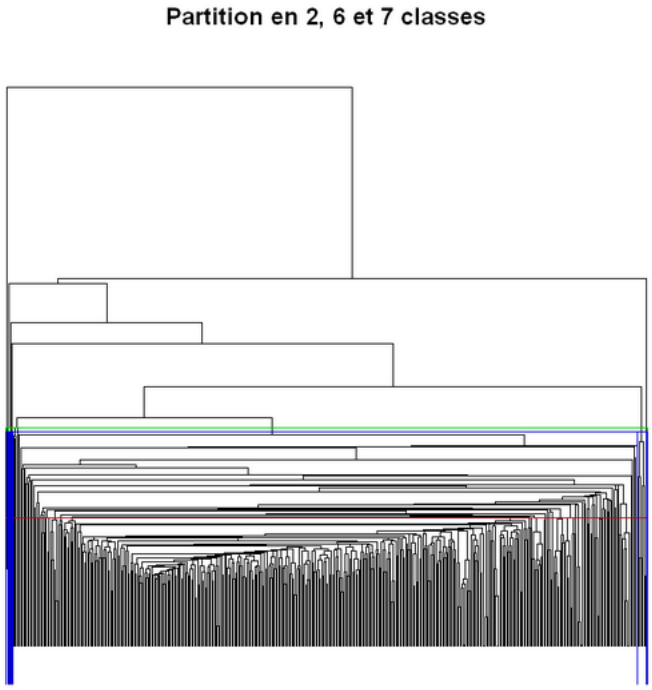


A première vue, on peut constater que l'arbre obtenu grâce à l'algorithme a du mal à extraire des classes distinctes de termes. Le dendrogramme n'est pas très lisible car les termes sont très nombreux.



Grâce au graphe de l'inertie totale, nous pouvons remarquer que le nombre de clusters qui maximise l'inertie sont 2,6 et 7

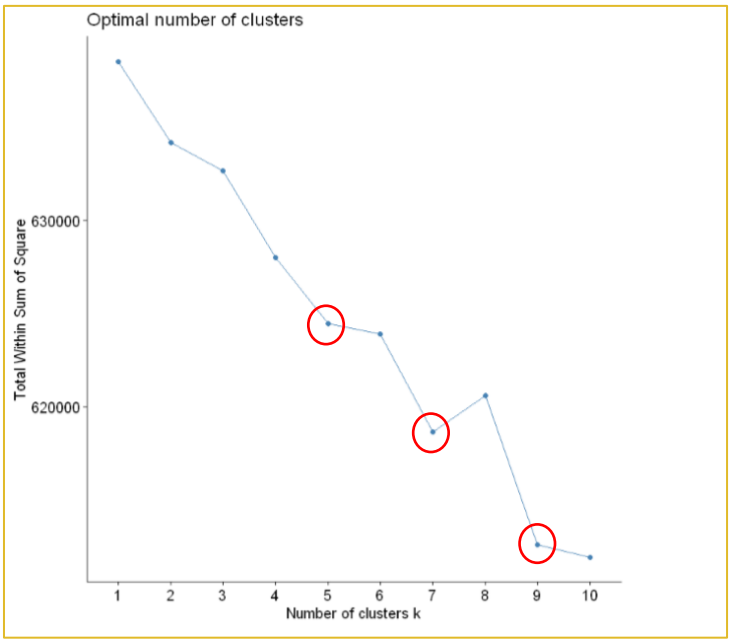
Dans le graphe ci-dessous, nous pouvons voir le résultat de ces trois options :

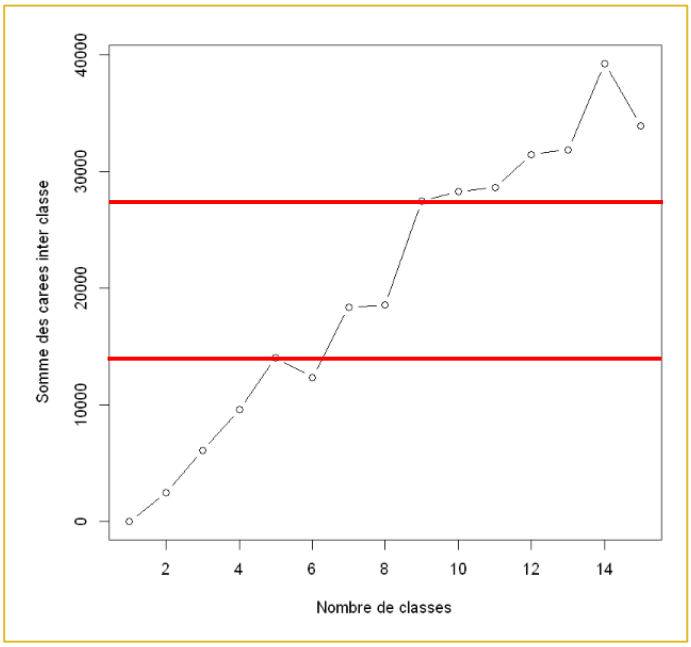


L'algorithme du CAH a mal repartitionné les termes, cela est dû à l'hétérogénéité des mots. En effet, les scores TF-IDF des termes sont assez rapprochés. La distance euclidienne entre les termes n'est donc pas assez grande pour partitionner les termes en cluster de manière optimale.

K-MEANS

Le K-Means est une technique largement utilisée en Data Science, elle permet de séparer les données en groupe, appelés des *clusters.* Nous proposons trois approches de K-Means dans ce rapport mais avant nous devons nous intéresser à l'évolution de l'inertie intra classe et inter classe. En effet, cette étude est primordiale pour déterminer un nombre de clusters optimal qui permette de minimiser l'inertie intra classe et maximiser l'inertie inter classe.

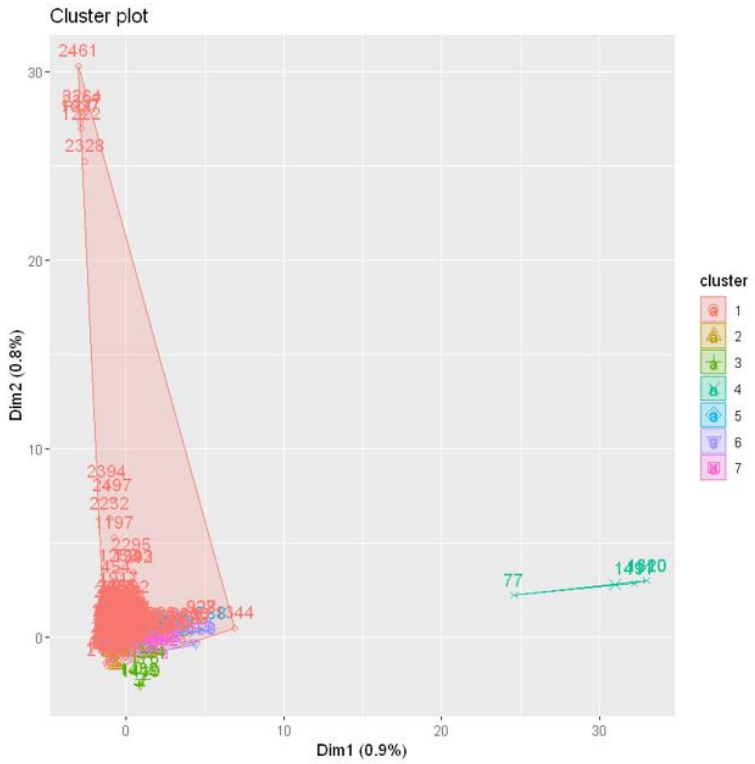




Sur le graphe de l'inertie inter classe, on remarque trois « coudes » à K = 5, 7 et 9. De plus, dans le second graphe on note que la courbe stagne à K=7. De plus, à partir de K=9 la courbe ne croit plus de manière significative.

De ce fait, le nombre de classe profitable qui maximise l'inertie inter classe tout en minimisant l'inertie intra classe est K = 7.

APPROCHE 1 : K-MEANS AVEC CENTRAGE ET REDUCTION DES DONNEES



INTERPRETATION DES CLUSTERS

Le graphe ci-dessus représente le clustering que nous avons obtenu avec la matrice de contingence une fois celle-ci centrée et réduite. Afin de se rendre compte de la pertinence des clusters produits, nous avons lu quelques tweets par clusters grâce à une matrice de passage qui conserve l'ID et le texte de chaque tweet.

CLUSTER 1

Ce cluster regroupe 1356 tweets. Sur trois tweets pris au hasard tout trois évoque le Rock'n'Roll.

CLUSTER 2

Ce cluster contient 3 tweets. Aucun n'évoque le même sujet ; l'un parle d'une certaine Nina Pop et d'une justice pour les « Black Lives », l'autre fait référence à Brandon Lewis, homme politique britanique, qui s'est exprimé à la BBC concernant le Covid-19 et le dernier est un tweet de deux mots « pop mint ». Ce cluster n'est malheureusement absolument pas informatif.

CLUSTER 3

Il regroupe 5 tweets dont la majorité est associé par l'univers cultural du reggae. Certains tweets parlent du style musical, d'autres évoquent la consommation de drogue et d'alcool. Néanmoins, on retrouve deux tweets qui parlent de rock'n'roll. De ce fait, ce cluster n'est pas correctement construit.

CLUSTER 4

Ce cluster regroupe 5 tweets. Aucun des tweets a quelques choses en commun puisque certains parle du Rock'n'Roll, d'autre de HipHop ou encore de Jazz.

CLUSTER 5 ET 6

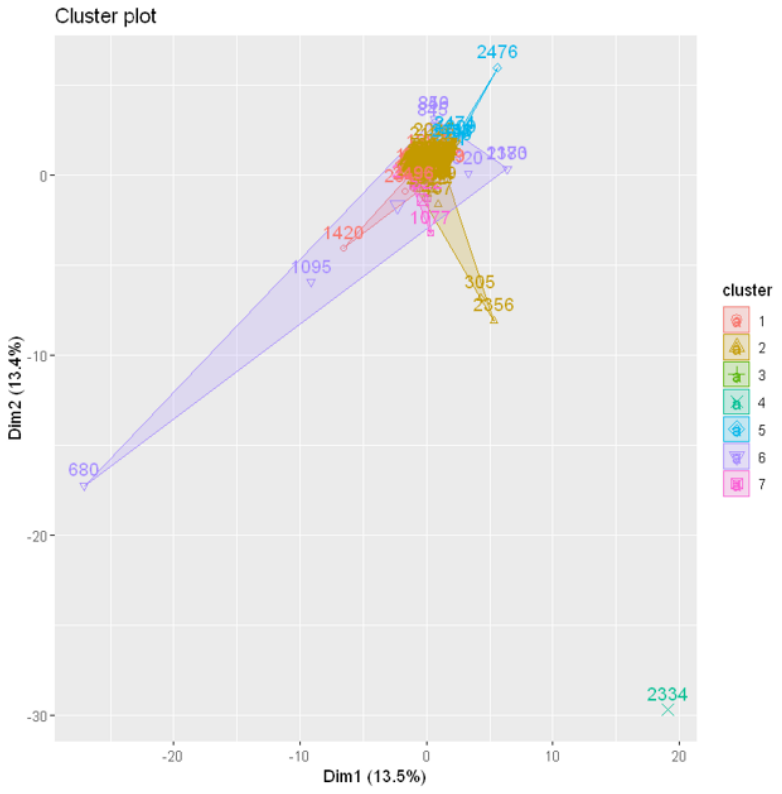
Ils contiennent respectivement 20 et 21 tweets. Dans ces clusters, la majorité des tweets évoque l'univers artistique du Jazz.

CLUSTER 7

Ce dernier cluster contient 10 tweets. Aucun style musical ne prend le dessus sur un autre dans ce groupe de tweet.

➔ Finalement, l'approche du K-Means par la matrice centrée réduite n'a pas apporté de résultats significatifs. C'est pourquoi il faut s'intéresser à l'approche du K-Means basée sur l'AFC.

APPROCHE 2 : K-MEANS SUR LES COMPOSANTES PRINCIPALES DE L'AFC



INTERPRETATION DES CLUSTERS

Ce graphe représente la répartition des tweets produite par l'approche du K-Means utilisant l'AFC.

CLUSTER 1

Dans ce cluster est regroupé 205 tweets. Sur les trois tweets pris dans la première approche du K-Means on en retrouve deux. Par ailleurs, le sujet commun de ces tweets est toujours majoritairement le Rock'n'Roll.

CLUSTER 2

Il regroupe 1164 tweets. Sur trois pris au hasard, tous évoquent le Rock'n'Roll.

CLUSTER 3

Ce cluster contient 2 tweets ; l'un évoque une radio de reggae, l'autre parle des bitcoins.

CLUSTER 4

Il contient seulement 1 tweet qui parle d'un orchestre d'instrument électrique qui va bientôt se produire.

CLUSTER 5

Ce cluster qui contient 8 tweets parle majoritairement de rap en évoquant des paroles de chanson, de la culture urbaine (graffitis)

CLUSTER 6

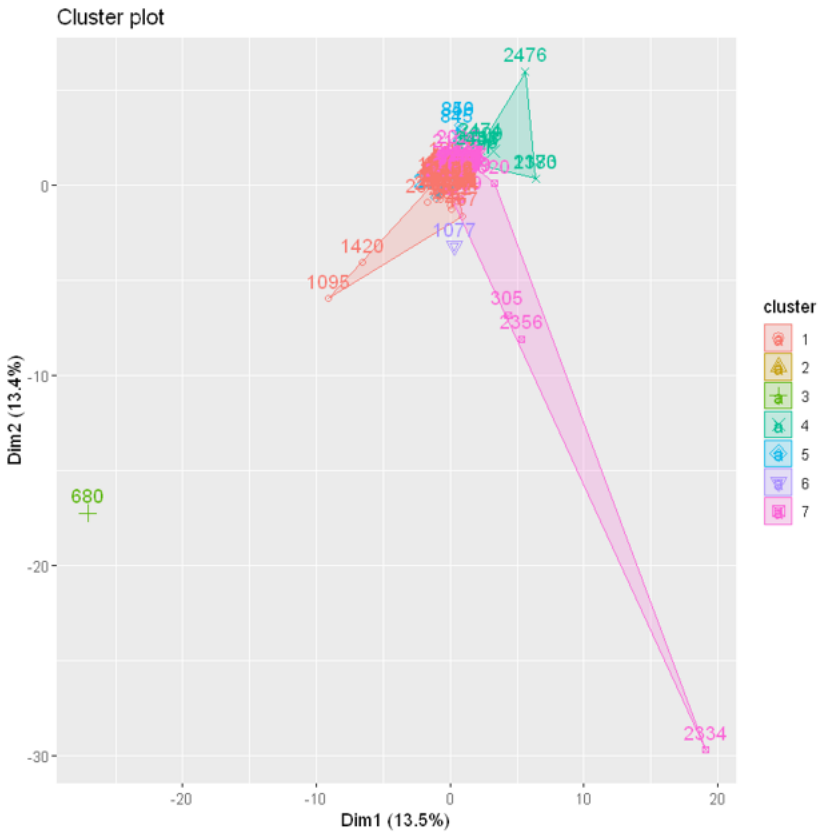
Aucun sujet commun ne ressort de ce cluster, on y compte 8 tweets certains parle de pop avec l'artiste Lady Gaga, d'autre du rappeur Kendrick Lamar ou encore du célèbre jeu Fortnite.

CLUSTER 7

Ce dernier cluster contient 4 tweets évoquant tous un style de musique particulier : rap, jazz, pop et blues.

➔ Une fois de plus, cette approche ne nous permet pas d'avoir des clusters nettement dissociés contenant des informations uniques.

APPROCHE 3 : K-MEANS SUR LES K-1 COMPOSANTES PRINCIPALES DE L'AFC



INTERPRETATION DES CLUSTERS

Ce graph représente les clusters construits à partir des K-1 composantes principales de l'AFC. Afin de se rendre compte de la pertinence de ces clusters, nous étudions au hasard quelques tweets issue de ces derniers.

CLUSTER 1

Ce groupe contient 1271 tweets. Sur cinq tweets pris au hasard, deux d'entre eux évoque des groupes de rock, les trois autres parlent du style de vie « rock'n'roll » ou associe le mot rock comme un adjectif à des plats culinaires.

CLUSTER 2

Ce cluster contient 2 tweets. Malheureusement, ces tweets n'ont pas un sujet commun, l'un évoque une parole de chanson de style musical Rap, l'autre fait mention d'une radio de Reggae.

CLUSTER 3

Ce cluster a qu'un seul tweet qui fait certainement référence à des paroles de chanson de Rap US.

CLUSTER 4

Ce cluster regroupe 10 tweets. Sur un panel de cinq tweets, tous parlent de musique mais de style music diffèrent comme le Rap, l'Electro ou encore les musiques des années 80.

CLUSTER 5

Il contient 5 tweets. Deux d'entre eux ne parle absolument pas de music l'un parle du Covid-19, l'autre demande justice pour les « Black Lives ». C'est deux tweets ont déjà été réunis dans un cluster au cours d'une autre approche de K-Means. Le reste des tweets parle de style music différents.

CLUSTER 6

Comme le cluster 3, il ne contient qu'un seul tweet. Celui évoque un chanteur de Rap connu Kendric Lamar.

CLUSTER 7

Ce cluster contient 102 tweets. Sur un panel de cinq tweet, trois d'entre eux parle du rock'n'roll. Un parle d'un jeune rappeur américain et le dernier parle d'une chanteuse italienne qui semble faire des clips controversés.

Finalement, ces trois approches sont assez compliquées à départager puisqu'aucune n'a su produire des clusters suffisamment séparés. Néanmoins, les approches utilisant l'AFC semblent plus réaliste que la première approche qui se sert de la matrice de contingence centrée réduite. Nous pensons que ce manque de distinction entre deux tweets différents est dû soit au trop grand nombre de termes employés dans la matrice de contingence soit dès la première étape c'est-à-dire l'extraction des tweets. Néanmoins, la méthode du K-Means semble être de toute évidence la meilleure technique de classification à employer du fait de la large base de données que nous utilisons.

CONCLUSION

Ce projet avait pour but d'extraire des tweets du réseau sociale afin d'appliquer de clusterings dessus.

Pour commencer nous avons extrait les tweets grâce à l'API public de Tweeter. Nous avons ensuite procédé au nettoyage des tweets afin d'en supprimer ceux qui sont redondants. Cela nous permet d'éviter d'avoir des résultats biaisés par la redondance. Une fois ces tweets nettoyés, on rassemble les termes qui composent ces tweets afin de créer un vocabulaire qui sera utilisé pour la matrice de contingence. Ce vocabulaire est également nettoyé grâce à différents procédés tel que la racinisation et la lemmatisation.

Dans un premier temps, nous avons effectué une AFC sur une matrice de contingence de fréquence des n-top termes par tweets. Cela nous a permis de visualiser la représentations des termes et des tweets ensemble et séparément. Dans un second temps, nous avons appliqué des algorithmes de cluster : CAH et K -means. Au vu des résultats obtenus, nous avons remarqué que les termes pris en compte ne sont pas des termes significatifs dans le domaine musical. Nous avons donc décidé de prendre les n top termes de matrice de contingence des scores TF-IDF. Ce score s'est avéré plus efficace qu'une simple fréquence. Toutefois, les n tops termes ne sont toujours pas aussi significatifs. Ce problème sera notamment remarqué lors du CAH, l'algorithme n'arrivant pas à partitionner les termes en différent clusters. De plus, le vocabulaire est hétérogène ce qui qui complique le procédé de minimisation de l'hétérogénéité intra classe. Le K-means combiné à l'AFC a toutefois obtenus des résultats assez satisfaisants. En effet, nous obtenons 7 clusters de tweets ce qui se rapproche de notre vérité. Ce que nous appelons vérité dans ce cas c'est l'extraction des 5 ensembles de tweets portant sur des thématiques musicales différentes. En ce qui concerne les clusters en plus, l'algorithme semble parfois détecter des similarité en plus du genre musicale.

Ce que nous pouvons conclure, des résultats de ce projet c'est que ni la fréquence ni le score TF-IDF ne sont assez efficaces pour représenter le liens document termes. Notamment dans notre cas, où la fréquence des termes par tweet ainsi que le nombre de termes par tweet ne sont pas assez grands.

En conclusion, grâce aux algorithmes de prédiction non supervisé vu en cours nous avons pu analyser les données et effectué des clusterings de données. Le contexte de ce projet, nous a également permit d'apprendre des techniques couramment utilisées dans le domaine du Data Mining, tel que le Bag of words ou encore le score du TF-IDF.