Evolution des lyrics de la musique Américaine

TEXT MINING

Nabil LOUDAOUI
Omar SAIP SY
Yann KIBAMBA
13/12/23 - 13/02/24
Université Rennnes 1
Encadré par Loic VERDIER

Introduction

Composé d'Omar SAIP, Nabil LOUDAOUI et Yann KBAMBA, notre a choisi comme sujet d'étude pour le projet de Text Mining : « L'Evolution des paroles de la musique américaine au fil du temps ». Notre choix s'est porté sur ce sujet pour plusieurs raisons. Tout d'abord, qui dit musique américaine dit musique anglophone qui est un langage universel et cette dernière parle à un large public. Ensuite, les paroles de chansons et les artistes constituent une source riche et diversifiée de mots et d'expressions. Enfin, l'aspect temporel de l'étude permettrait de mettre en lumière les caractéristiques typiques des différentes époques à travers l'évolution des thèmes, des styles et des messages véhiculés dans les paroles de chansons.

En somme, l'objectif est de comparer ce domaine suivant différents axes majeurs pour se faire une idée générale du niveau lexicale, de l'engament des artistes etc…

Pour mener à bien notre analyse, nous avons fait le choix d'utiliser le logiciel Python et ses librairies spécialisées, afin de réaliser nos tâches liées au traitement de texte ou encore au scraping des données et à la transformation de la donnée.

Ce rapport a pour but de présenter de manière détaillée les différentes étapes que nous avons suivies pour réaliser notre projet de Text Mining. Il se structurera en trois parties principales. Dans un premier volet, nous aborderons la recherche et l'acquisition de la source de données. Ensuite, nous détaillerons les différentes analyses descriptives et traitements que nous avons réalisés. Enfin, nous présenterons les résultats obtenus avant de conclure en portant un regard sur les limites et ouvrant la voie à des pistes pour de futures recherches dans ce domaine.

Partie 1 : Constitution de la base de données

Pour constituer notre corpus et obtenir une représentation fidèle de la musique américaine, nous avons choisi de nous concentrer sur trois styles musicaux fondamentaux des trois dernières décennies (1990-2000, 2000-2010 et 2010-2020) : le rap, la pop et le rock. Afin d'assurer la représentativité de nos résultats, nous avons inclus les sept artistes les plus influents (à titre subjectif) de chaque style pour chaque époque, avec 60 chansons pour chaque artiste. Cela nous donne un jeu de données comprenant 63 artistes et 3780 chansons. Il est à noter que certains artistes se sont illustrés sur deux ou trois décennies. Dans ces cas, nous avons pris soin de les inclure à l'époque où leur notoriété était la plus importante, en nous basant sur nos connaissances et des articles spécialisés.

Après avoir défini le cadre de notre base de données, il a fallu trouver une source de données fiable. Nous avons opté pour le site "Paroles de Chansons : Paroles et traductions de vos chansons préférées" sur lemonde.fr, qui publie un grand nombre de paroles de chansons. Pour récupérer les données, le scraping s'est avéré être la meilleure solution pour automatiser le processus.

Dans un premier temps, nous avons analysé la structure du site et le code source correspondant. Une fois cette étape franchie, nous avons utilisé un vecteur contenant les noms des artistes souhaités pour extraire les données nécessaires. Cette étape s'est révélée chronophage (environ 1 heure), c'est pourquoi nous avons exporté notre base de données au format CSV pour faciliter le traitement ultérieur.

Pendant la constitution de notre base de données, nous avons rencontré plusieurs problématiques :

• Les chansons étaient réparties sur plusieurs pages, avec un nombre variable de chansons par page pour chaque artiste, rendant la récupération de données uniforme assez complexe.

• N'ayant pas de moyen de savoir quelles chansons étaient ajoutées, car certaines étaient des remixes avec une sémantique commune, et des collaborations entre les artistes sélectionnés ont conduit à des doublons.

• Le principal défi a été de traiter les titres de chansons pour l'exportation. En effet, les titres apparaissent dans les URL avec des tirets à la place des espaces, mais les ponctuations et certains symboles comme le dollar américain ont une représentation différente dans les URL.

Avant toute analyse, nous avons dû nettoyer notre base de données pour nous assurer de sa conformité.

Ci-dessous 15 lignes aléatoires de notre base de données :

	Artiste	Sexe	Style	Decennie	Chanson	Paroles
3187	foo-fighters	Homme	Rock	2000s	for-all-the-cows	I'm called a cowI'm not aboutTo blow it nowFo...
2628	pearl-jam	Homme	Rock	1990s	get-it-back	On the side of the roadLost and aloneTo get i...
168	nas	Homme	Rap/Hip-Hop	1990s	cruise-control	[Joey Bada$$]Livin' in a world so coldKeep a ...
871	drake	Homme	Rap/Hip-Hop	2010s	big-amount	[2 Chainz]I'ma tell you, I'ma tell you this r...
1851	alicia-keys	Femme	Pop	2000s	intro-scenic-drive	[Khalid]Can we just talk? Can we just talk?Se...
1385	janet-jackson	Femme	Pop	1990s	made-for-now	[Daddy Yankee]Janet JacksonIconicDaddy[Janet ...
2760	radiohead	Homme	Rock	1990s	creep	When you were here beforeCouldn't look you in...
322	ice-cube	Homme	Rap/Hip-Hop	1990s	damn-homie	[50 Cent]Damn, homieIn high school you was th...
2939	u2	Homme	Rock	1990s	lemon	LemonSee through in the sunlightShe wore lemo...
2976	linkin-park	Homme	Rock	2000s	heavy	I don't like my mind right nowStacking up pro...
1785	rihanna	Femme	Pop	2000s	half-of-me	You saw me on a televisionSetting fire to all...
1152	nicki-minaj	Femme	Rap/Hip-Hop	2010s	everybody	[Nicki Minaj]MmAin't gang if you let shit sli...
1019	kendrick-lamar	Homme	Rap/Hip-Hop	2010s	hood-politics	K dot, pick up the phone, every time I call i...
2372	katy-perry	Femme	Pop	2010s	into-me-you-see	I built a wall so high, no one could reachA l...
2592	pearl-jam	Homme	Rock	1990s	leatherman	I know about a man to whom I may be related, ...

Notre base de données est répartit de la manière suivante :

• 1177 chansons Pop, 1109 de Rap/Hip-Hop et 1089 de Rock

• 1157 1990s, 1131 2000s et 1087 2010s

• 2388 chansons d'hommes, 936 de femmes et 58 de groupe (mixte)

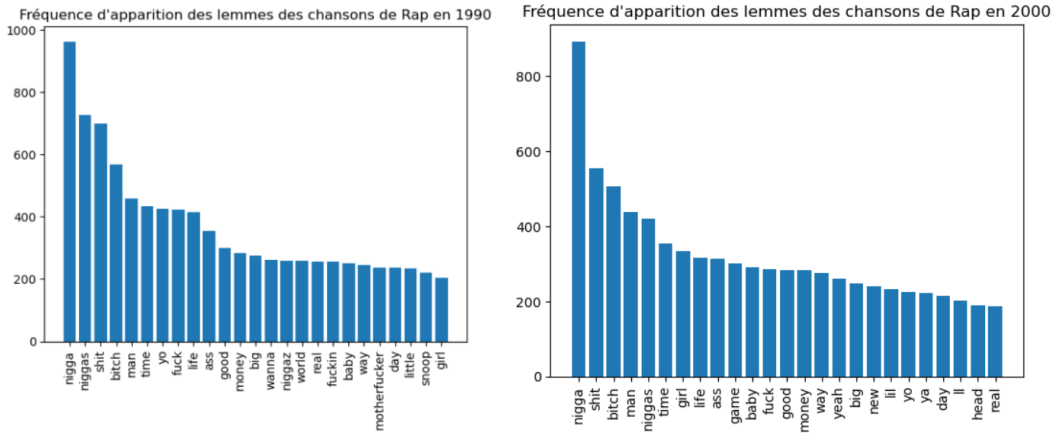
Nos axes d'études sont équilibrés.

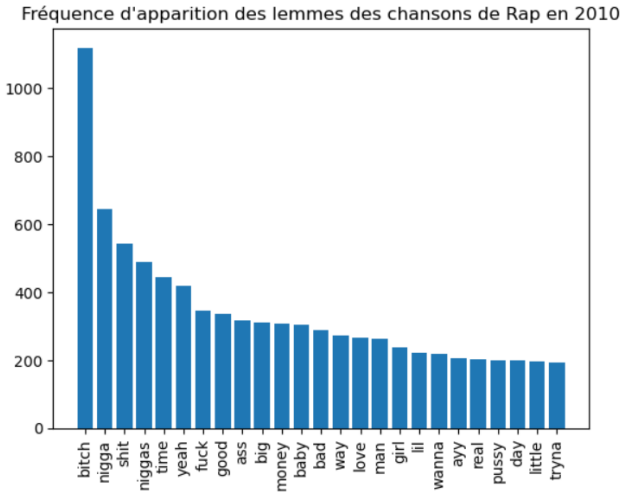
En résumé, pour cette section, il aurait été intéressant d'envisager la comparaison avec d'autres plateformes de sources de données telles que "Genius" pour optimiser la qualité de la base de données et réduire le temps de traitement. De plus, nous aurions pu explorer l'existence d'API pour faciliter la récupération des données.

Partie 2 : Analyse descriptive

Pour le processus de nettoyage sémantique, nous avons commencé par retirer les Stops Word, puis nous avons observé graphiquement les tendances émergentes de nos corpus.

Pour le Rap/Hip-Hop :





Sur les trois décennies, nous observons une prévalence presque identique des mêmes termes. A première vue, on remarque l'emploi d'un champ lexical familier, d'argot et d'interjection. L'analyse des barplots a permis de mettre en lumière des variations et des évolutions dans le langage dans le langage courant comme celui de l'argot. Par exemple, on va retrouver la présence de « niggas » et « niggaz » pour la version plurielle de « nigga ». Par la suite, ces deux termes seront regroupés en un seul.

Les premières conclusions que nous pouvons tirer sont les suivantes :

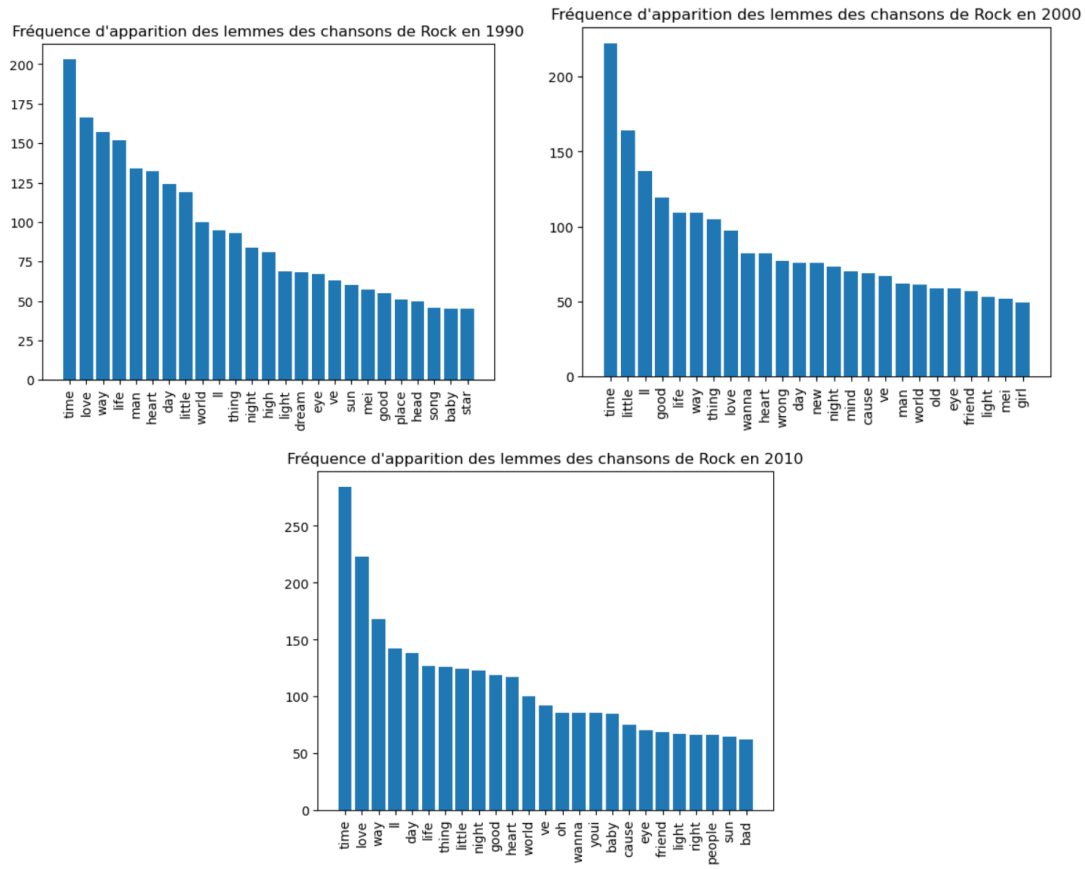
• L'emploi fréquent d'un langage vulgaire et cru, dans un but de renforcer l'authenticité de l'artiste va conférer une dureté et une intensité particulière aux paroles.

• Exploration récurrente des thèmes de genres, de relations et de dynamiques sociales

• Références à la richesse et au statut social, suggérant une préoccupation aux aspects économiques.

• Il est notable que le champ lexical des armes n'est pas aussi prédominant que prévu

Pour le rock :



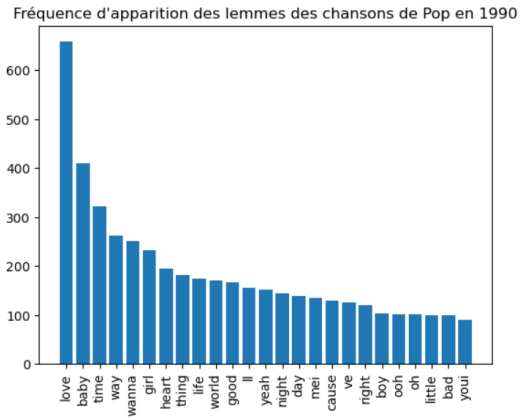
A travers les trois décennies, une certaine continuité se dégage dans les choix de mots choisis. Les termes « Time », « love » ou encore « way » laisse penser à des préoccupations humaines de la vie et des thèmes universels.

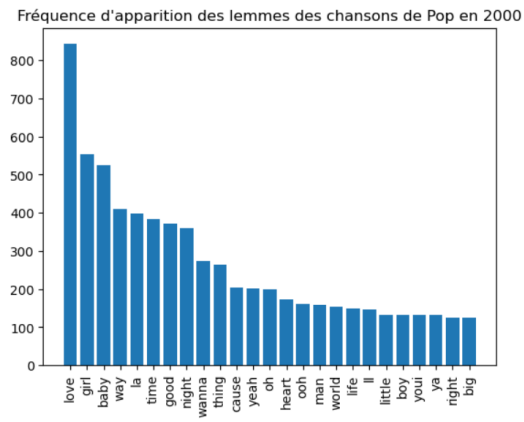
Pourtant, des variations intéressantes se manisfestent. Dans les années 90 il y a une emphase sur des mots comme « dream », « high », « star » évoquant une énergie d'élévation et optimiste. Les années 2000 introduisent le terme « wrong », tandis que les années 2010 voient l'émergence du terme « bad ».

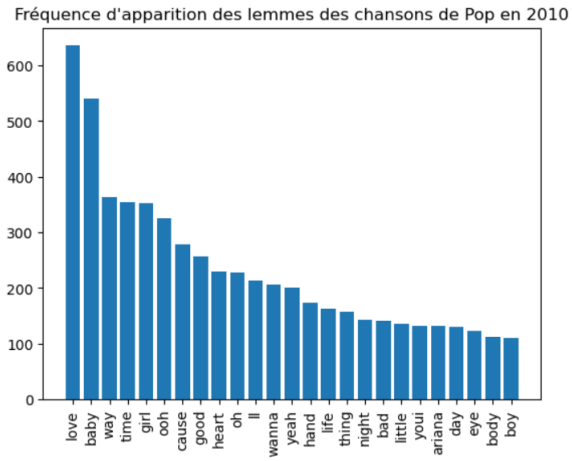
Ces termes pourraient indiquer une sensibilité plus consciente ou encore l'exploration de thèmes de confusion ou de frustration.

En somme, comme première conclusion, nous dirions que l'aspect social et les relations humaines sont au centre des paroles des artistes de Rock.

Pour la Pop :







Les termes « love », « baby », « way » ou encore « girl » fortement représentés dans les trois décennies, reflètent une prédominance des thèmes en lien avec l'amour et les relations interpersonnelles. On note qu'à partir des années 2000, l'apparition du terme « boy » pouvant suggérer l'apparition d'une perspective masculine en plus des thèmes traditionnels centrés sur les femmes. Quant aux années 2010, les termes « body » et eye » indique une évolution vers un champ lexical plus axé sur le physique, corporelle et visuelle.

En somme, tout comme les deux autres styles musicaux, la pop semble également s'inscrire au fil du temps avec des sujets liés à l'amour ou encore à l'expression de soi.

Partie 3 : Résultats

Pour approfondir l'analyse des paroles employées par les artistes, nous avons entrepris de classer les artistes afin d'évaluer leur richesse lexicale. L'objectif est d'obtenir le nombre de mots utilisés par artistes pondéré par son nombre de chanson présente dans la base de données. Pour ce faire, nous avons utilisé le dictionnaire de mots anglais de la librairie NLTK pour filtrer toutes les paroles n'y figurant pas comme l'argot. Puis nous avons établis notre indicateur. Le résultat présenté ci-dessous est agrégé par la décennie et le style musical :

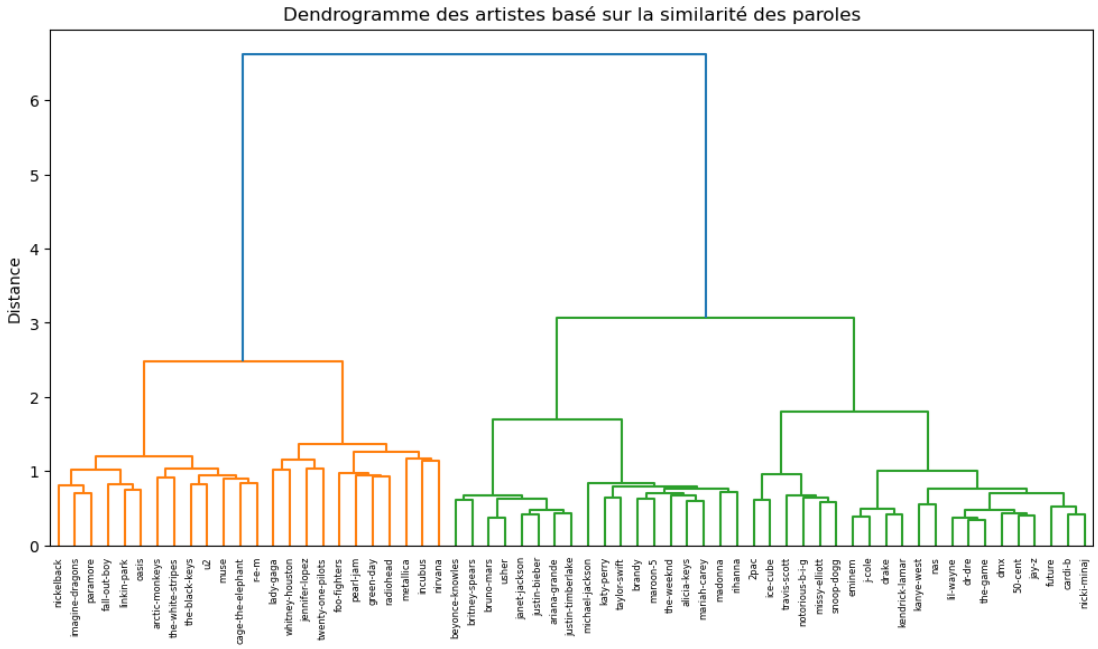
	Decennie	Style	mots_uniques_par_chanson
0	1990s	Pop	16.677196
1	1990s	Rap/Hip-Hop	34.597179
2	1990s	Rock	14.769488
3	2000s	Pop	18.604193
4	2000s	Rap/Hip-Hop	37.190560
5	2000s	Rock	17.285857
6	2010s	Pop	16.533780
7	2010s	Rap/Hip-Hop	29.591851
8	2010s	Rock	16.262689

• Dans le classement par artiste, tous les artistes de Rap/Hip-Hop occupent les premières places.

• On remarque que peu importe la décennie, la moyenne pour la catégorie Rap/Hip-Hop est nettement supérieur. Pour cette catégorie, on note par ailleurs une nette chute pour la dernière décennie.

• Il semblerait que les artistes Pop utilisent très légèrement plus de termes que les artistes de Rock.

Par la suite, nous nous sommes intéressés à la classification des artistes. Dans notre base de données ils sont répartis selon des critères prédéfinis comme le style musical (qui ne couvre pas nécessairement l'ensemble des chansons d'un artiste) ou encore la décennie, impliquant des époques et perceptions propres. En utilisant les paroles des artistes nous avons créé un dendrogramme pour observer la proximité entre les artistes.

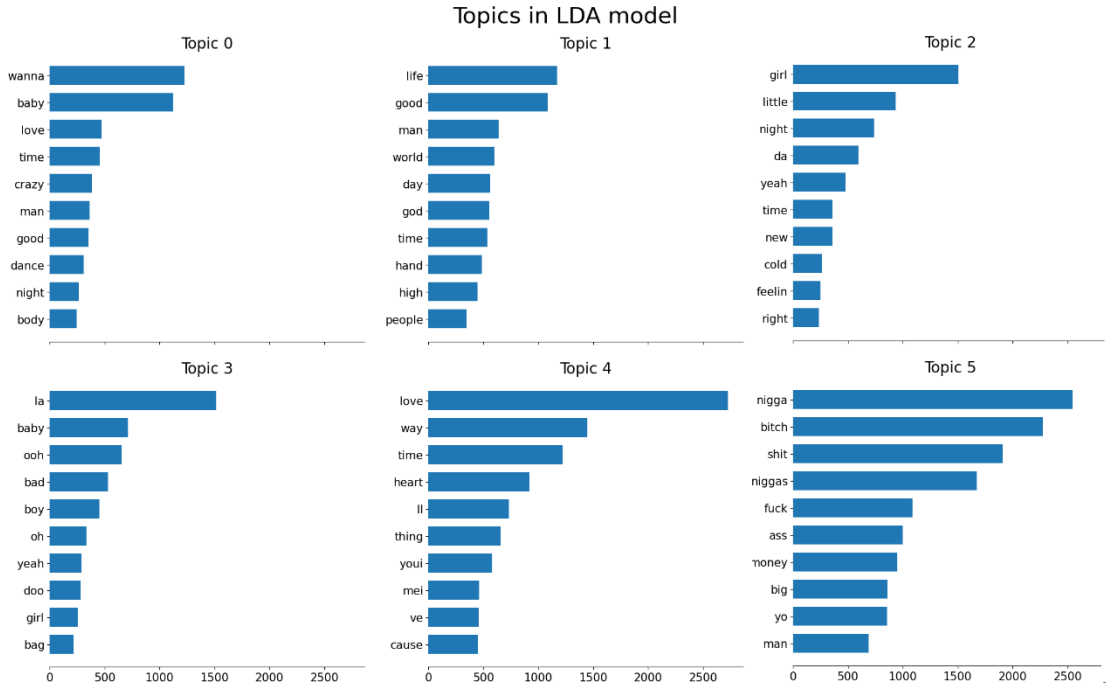


Le dendrogramme se rapproche très fortement de la réalité. Les deux premières branches montrent clairement la séparation entre le rock/pop et le rap/pop. A l'extrême à gauche nous observons les groupes de rock pur tandis que sur la seconde section nous retrouverons des groupes de rock mêlés à des artistes Pop comme Jennifer Lopez ou encore Lady Gaga. Du côté droit, la distinction entre les artistes de pop et de rap est nette. La proximité de Nicki Minaj er de Cardi B ou encore celle de Taylor Swifft et de Katy est très réaliste tant sur le plan musical que temporel.

Par la suite, nous avons cherchés à aller plus loin en utilisant les lyrics pour définir de nouvelles classes et ainsi obtenir une nouvelle classification à la fois des artistes et des styles en nous focalisant sur les thématiques abordées.

Nous avons pris l'ensemble des paroles de notre base de données, puis avons appliqué un Latent Dirichlet Analysis afin de faire ressortir les 6 thématiques les plus présentes.

De ce fait, nous avons obtenu le graphe suivant :

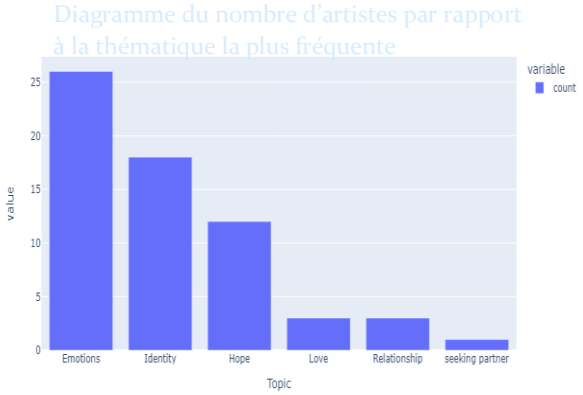


En partant de ce graphe, nous avons nommé les différentes thématiques par notre propre imagination (en fonction des mots qui composent la thématique). Nous avons :

- Topic 0: Relationship
- Topic 1: Hope (Motivation)
- Topic 2: Love
- Topic 3: Seeking partner
- Topic 4: Emotions
- Topic 5: Identity

Ensuite, nous avons cherché pour chaque artiste (et chaque style musical et chaque décennie) la thématique la plus fréquente parmi les 6 thématiques.

Nous obtenons les résultats suivants :



Topic	Artiste
Emotions	['britney-spears', 'madonna', 'janet-jackson', 'whitney-houston', 'michael-jackson', 'brandy', 'mariah-carey', 'rihanna', 'alicia-keys', 'justin-timberlake', 'jennifer-lopez', 'maroon-5', 'usher', 'taylor-swift', 'katy-perry', 'the-weeknd', 'justin-bieber', 'oasis', 'u2', 'linkin-park', 'muse', 'nickelback', 'imagine-dragons', 'the-black-keys', 'paramore', 'fall-out-boy']
Hope	['eminem', 'kanye-west', 'nirvana', 'pearl-jam', 'r-e-m', 'metallica', 'radiohead', 'green-day', 'foo-fighters', 'incubus', 'twenty-one-pilots', 'arctic-monkeys']
Identity	['2pac', 'dr-dre', 'nas', 'snoop-dogg', 'notorious-b-i-g', 'ice-cube', 'dmx', '50-cent', 'jay-z', 'lil-wayne', 'missy-elliott', 'the-game', 'drake', 'kendrick-lamar', 'cardi-b', 'j-cole', 'nicki-minaj', 'future']
Love	['travis-scott', 'the-white-stripes', 'cage-the-elephant']
Relationship	['beyonce-knowles', 'bruno-mars', 'lady-gaga']
seeking partner	['ariana-grande']

On remarque sur le diagramme que, parmi les 63 artistes de notre base de données, 26 abordent la thématique « Emotions » et 18 celle « Identity ». Ce qui recouvrent 2/3 des artistes environ. De ce fait on aurait pu faire ressortir 3 thématiques plutôt que 6 vu que les 3 dernières ne sont pas si fréquentes. Cependant, la diminution du nombre de thématique pourrait entrainer la perte d'une autre (par exemple en passant de 6 thématiques à 3, on perd celle « Hope »)

On peut aussi voir, à côté du diagramme la liste des artistes et leur thématique la plus fréquente. On note quelques faits choquant comme Travis Scott qui aborde plus la thématique « Love » plutôt que « Identity » ainsi que Linkin Park qui abordent « Emotions » plutôt que « Hope ».

Concernant le style musical et la décennie, nous obtenons les résultats suivants :

Style	Decade	Most frequent topic
Rap	1990s	Identity
	2000s	Identity
	2010s	Identity
Rock	1990s	Hope
	2000s	Emotions
	2010s	Emotions
Pop	1990s	Emotions
	2000s	Emotions
	2010s	Emotions

On peut voir qu'il n'y a pratiquement eu aucune évolution de thématique pour les 3 styles musicaux ; à part pour le Rock qui est passe de « Hope » à « Emotions ».

Conclusion

Cette étude soulève de nombreuses interrogations. L'exploration à travers trois décennies laisse penser que les époques ont forcément un impact respectif sur la musique. Notamment avec l'avènement progressif des réseaux sociaux ou les améliorations significatives des réseaux mobiles ont incontestablement façonné le paysage musical. Cependant, au vue des mots qui ressortent le plus nous serions tentés de dire que les thématiques n'ont pas changé sur les 30 dernières années.

Nous avions mentionné précédemment la présence de termes plus corporels dans le style musical de la Pop, ce qui suggère une liaison étroite avec la connexion des consommateurs devant un écran. De plus, pour étayer nos constats, il serait pertinent d'inclure des articles de presse ou des avis sur les réseaux sociaux pour confronter les paroles des artistes à l'opinion publique.

Bien que nos analyses de sentiment et de N-grams n'aient pas été concluantes, nous retenons du projet l'importance d'approfondir à la fois le travail sur les corpus mais également la compréhension sur la sémantique de la langue à l'étude.

Par soucis de temps, nous n'avons pas pu répondu à toutes les questions auxquelles nous aurions aimé répondre, cependant ce projet ouvre la voie à plusieurs perspectives de recherche intéressantes. Parmi celles-ci, nous pourrions explorer l'impact des paroles de chansons sur la société, ou sur des segments spécifiques de celle-ci. D'autres sujets d'étude pourraient inclure l'analyse plus précise de l'évolution des thèmes musicaux en incluant bien d'autres styles musicaux, l'impact de la musique sur les émotions et le bien-être, ou encore l'engament des artistes dans les sujets de sociétés.