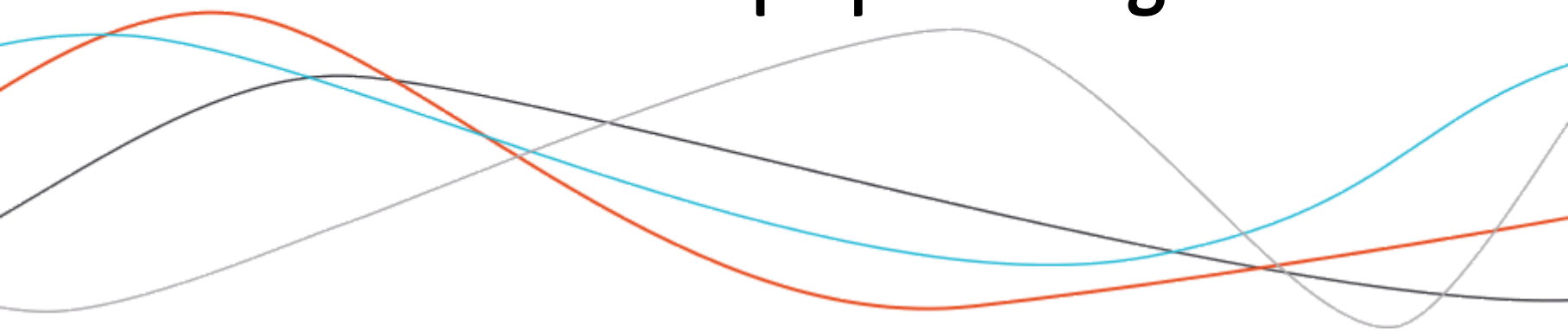
Le text-mining

Expliqué à ma grand-mère



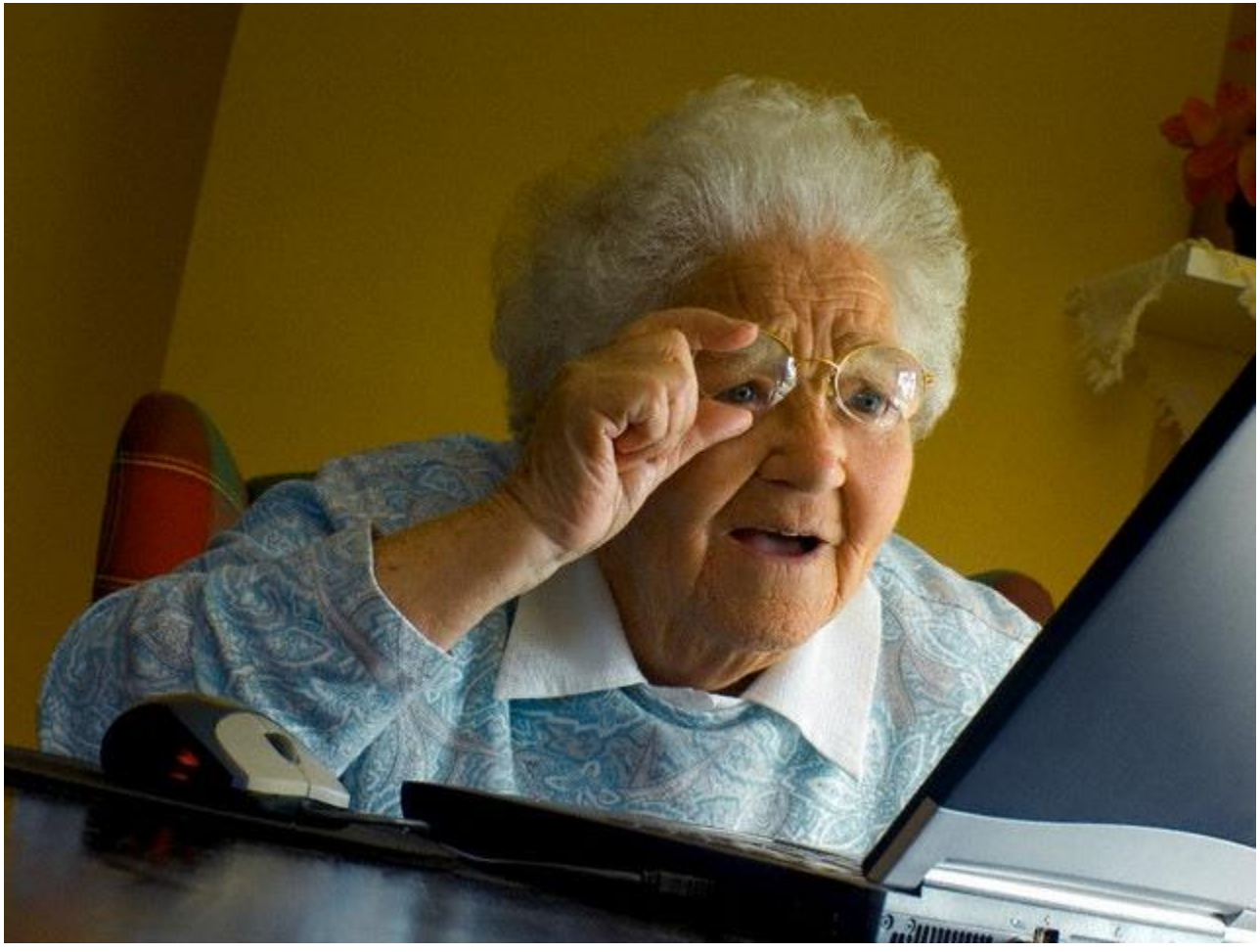
$ whoami

Colin FAY

Data Analyst, formateur R, expert Social Media à ThinkR, une agence spécialisée en Data Science et en langage R.

- http://thinkr.fr
- http://twitter.com/thinkr_fr
- http://twitter.com/_colinfay
- http://github.com/colinfay

Le text-mining expliqué à ma grand mère



C'est quoi le "texte mi-ninje" ?

Lexicométrie

> *(Linguistique) Science qui étudie statistiquement l'usage des mots.*

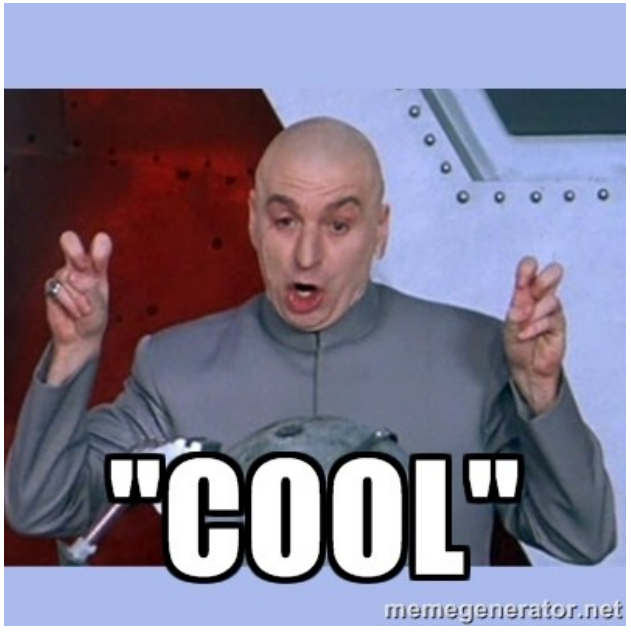
https://fr.wiktionary.org/wiki/lexicométrie

En clair, on transforme des mots en unités comptable.

> *Comme le métier ?*

Comme l'usage statisique. On parle aussi de textométrie, de traitement du language naturel, ou de forage de texte.

Mais puisqu'on est **jeunes et cools**, on utilise "text-mining" ou "natural language processing".



Text mining ou Natural language processing ?

Text mining

On parle de text mining quand on travaille plutôt du côté de la lexicometrie, c'est-à-dire quand on cherche à expliquer statistiquement un texte.

En text-mining, on va rechercher des tendances, décrire un texte par des chiffres. On est du côté des méthodes de "word frequencies".

Natural Language Processing (NLP)

Le traitement de langue naturel est quant à lui un discipline plus large, c'est-à-dire englobant le text-mining mais aussi d'autres méthodes de machine learning.

Par exemple, le topic modeling, la création de cluster, la prediction, etc.

> *Ça m'a l'air bien compliqué*

Dans la pratique, on utilise ces deux termes indistinctement.

Dans la vraie vie, ça sert à quoi ?

Il y a énormément de cas d'usage du traitement du langage naturel.

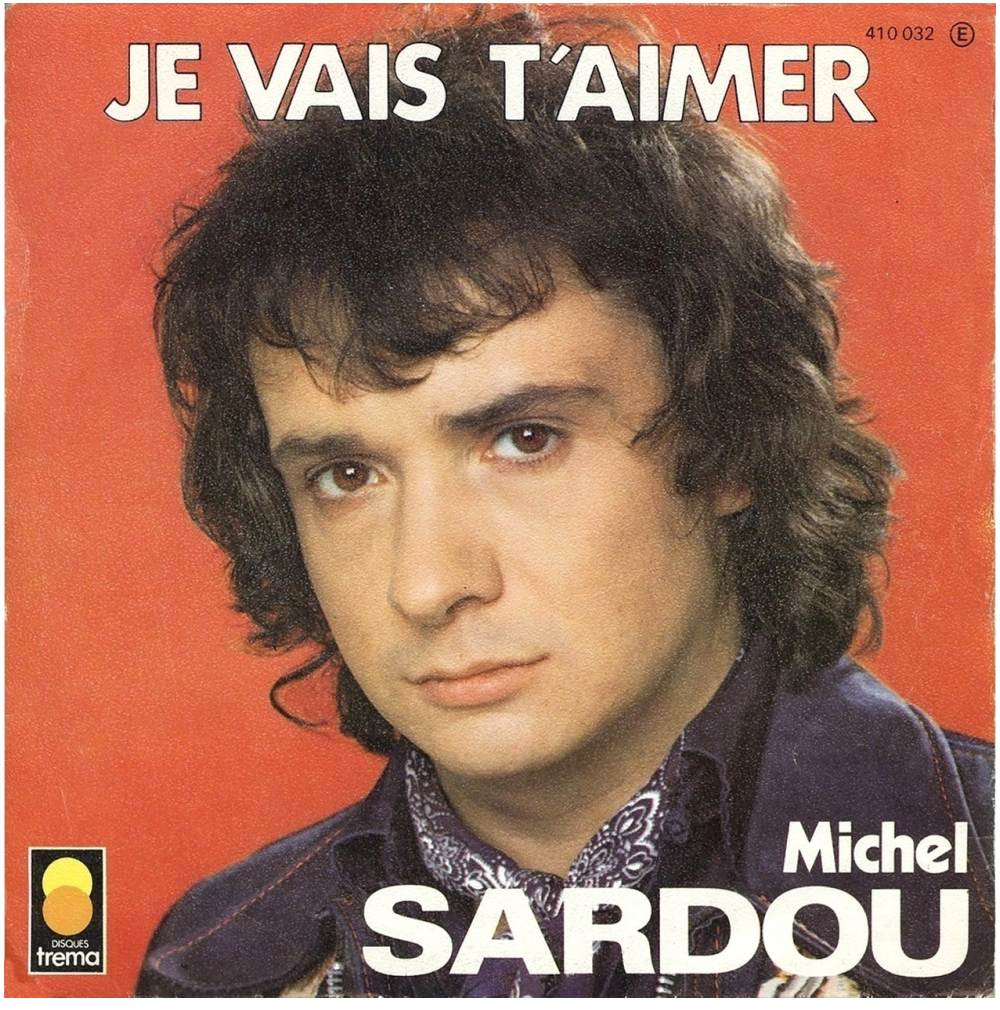
On peut notamment penser à :

- La détection de SPAM
- La traduction
- Les assistants type Siri ou Google Home
- Le marketing
- La relation client
- Les moteurs de recherche

etc...

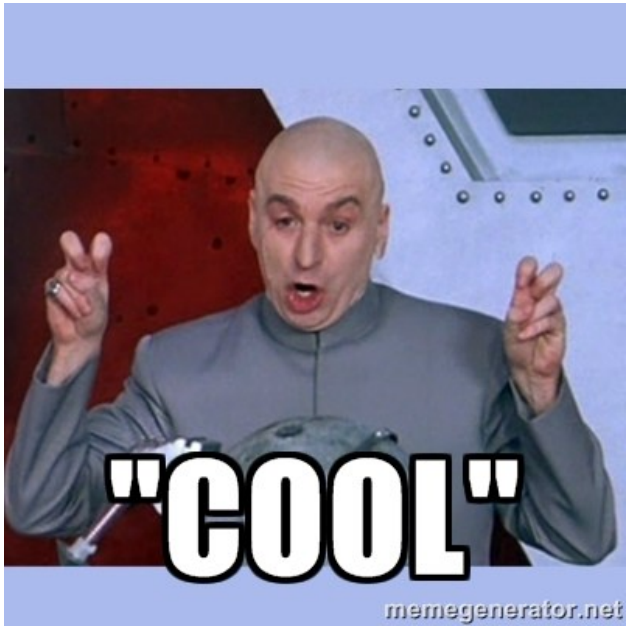
> *... Oui enfin ça c'est ta vie à toi, pas la mienne !*

Pour toi Mamie



Commençons par le commencement

La première approche, et peut-être la plus intuitive, est l'approche dite 'bag of words', ou 'sac de mots', mais **on utilise le terme anglais si on veut avoir l'air cool**.

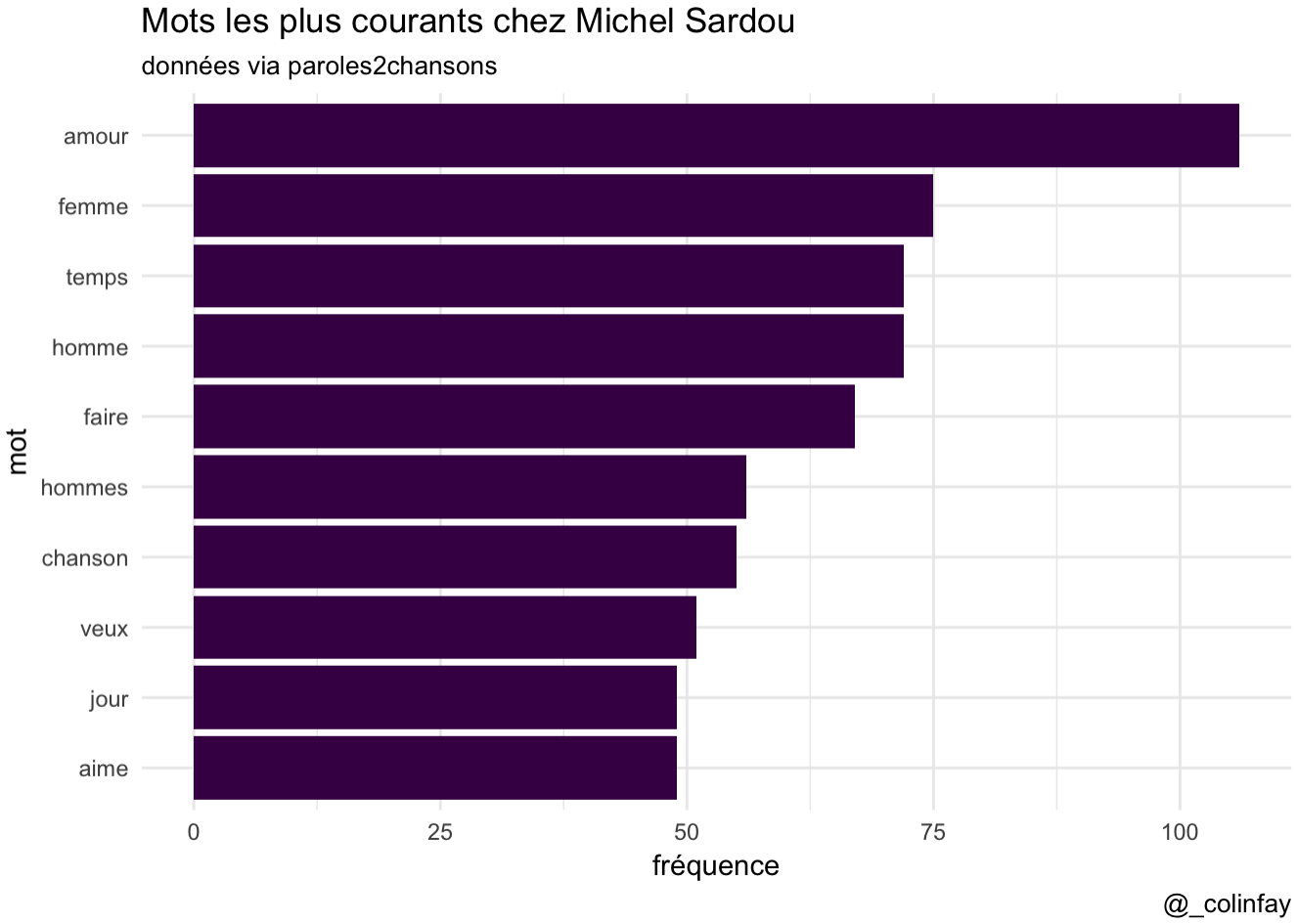


En clair, on imagine qu'on sépare tout le texte en unités, qu'on les mets dans un grand sac ensemble, et qu'on compte le nombre d'occurrences de chaque unité.

Autrement dit, on répond à la question "quels sont les termes les plus utilisés ?"

Un exemple ?

De quoi parle Sardou ?



De quoi parle Sardou ?

Ici, on imagine que nous avons mis tous les mots de toutes les chansons de Sardou, que nous les avons tous compté un par un, en notant la fréquence de chaque. Résultat ? Le terme qui ressort le plus est "amour".

> *"Attends, ça devrait plutôt être "le", non ?*

Bonne remarque, en text-mining, nous avons une méthode qui **consiste à retirer les mots vides de sens du corpus**, que l'on appelle également "stopwords".

Par exemple, en français, on retrouve :

```
#>   [1] a              abord          absolument
#>   [4] afin           ah             ai
#>   [7] aie            aient          aies
#>  [10] ailleurs       ainsi          ait
#>  [13] allaient       allo           allons
#>  [16] allô           alors          anterieur
#>  [19] anterieure     anterieures    apres
#>  [22] après          as             assez
#>  [25] attendu        au             aucun
#>  [28] aucune         aucuns         aujourd
#>  [31] aujourd'hui    aupres         auquel
#>  [34] aura           aurai          auraient
```

De quoi parle Sardou ?

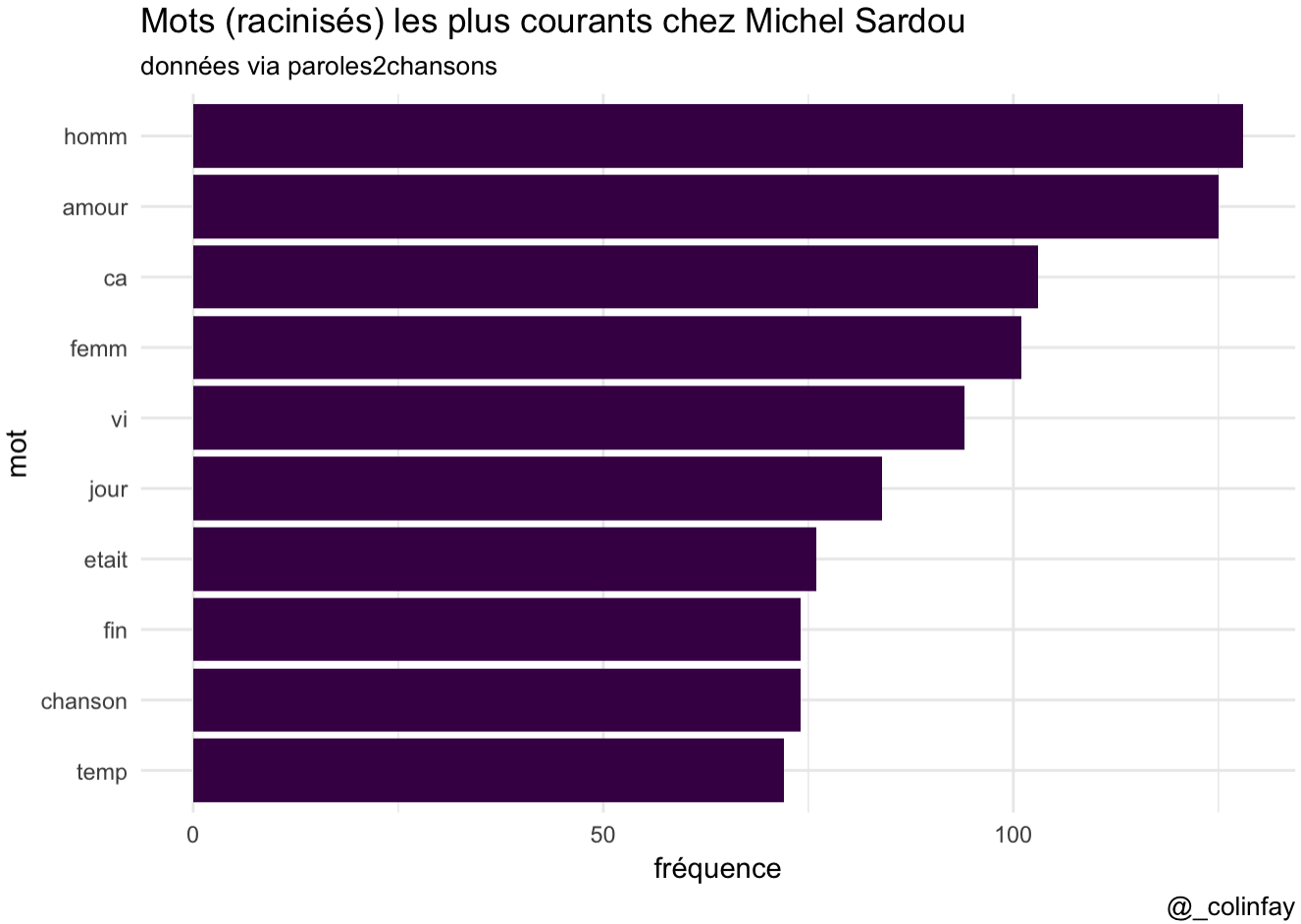
> Bon ce qui m'embête aussi, c'est qu'on retrouve "homme" et "hommes", c'est le même sujet non ?

Effectivement, pour cela on va faire appel à une méthode de lemmatisation ou de racinisation.

La différence étant que la première regroupe vers "la plus petite unité commune de sens", la deuxième prend "juste" les racines des mots.

Pas de bonne réponse sur quelle méthode est la meilleure, il faut tester.

De quoi parle Sardou ?



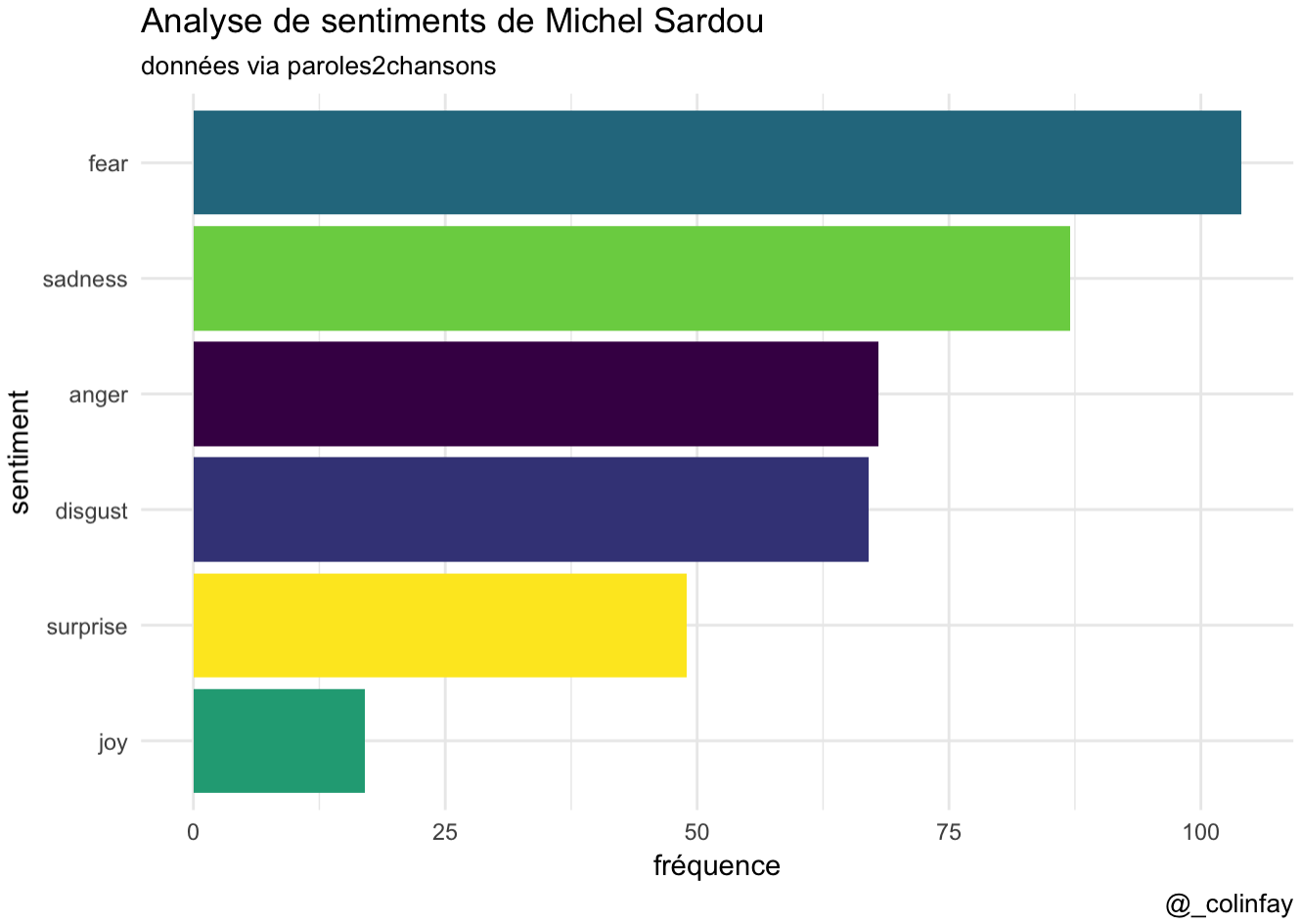
Sentiment analysis

Toujours parce qu'on est super cools, on utilise ensuite des méthodes de "sentiment analysis", c'est-à-dire une extraction automatique des sentiments liés aux mots.

Pour ça, on utilise des dictionnaires préfaits, parce que bon, tout faire à la main ça prendrait du temps.

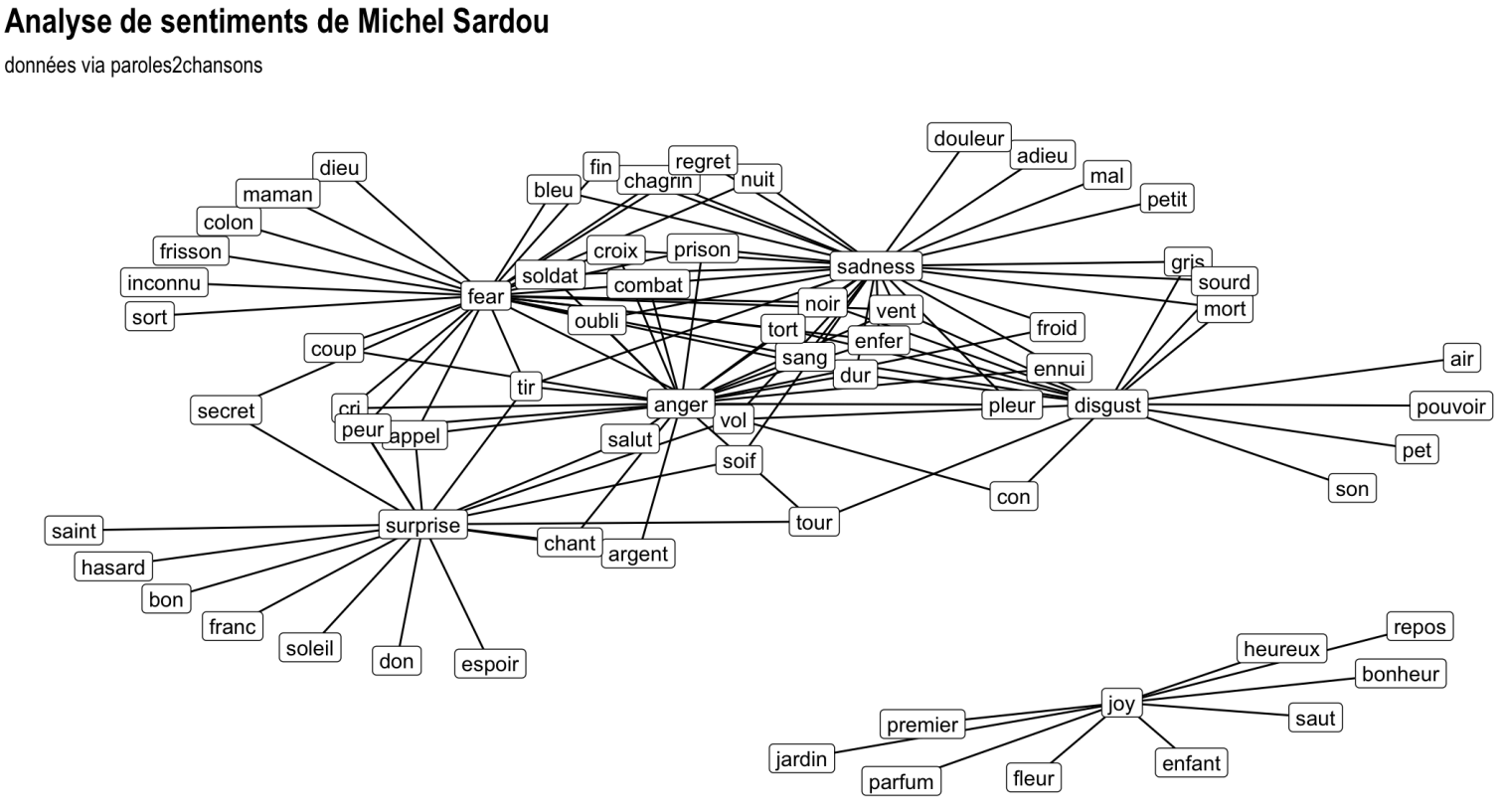
```
#> # A tibble: 10 x 2
#>            word sentiment
#>           <chr>     <chr>
#>  1 abaissement   sadness
#>  2    abaisser   sadness
#>  3     abandon      fear
#>  4     abandon   sadness
#>  5     abandon     anger
#>  6     abandon  surprise
#>  7  abandonner      fear
#>  8  abandonner   sadness
#>  9  abandonner     anger
#> 10  abandonner   disgust
```

Sentiment analysis



Sentiment analysis

Quand on est cool, on essaie de visualiser un text par d'autres moyens qu'un "barchart"



À travers les âges

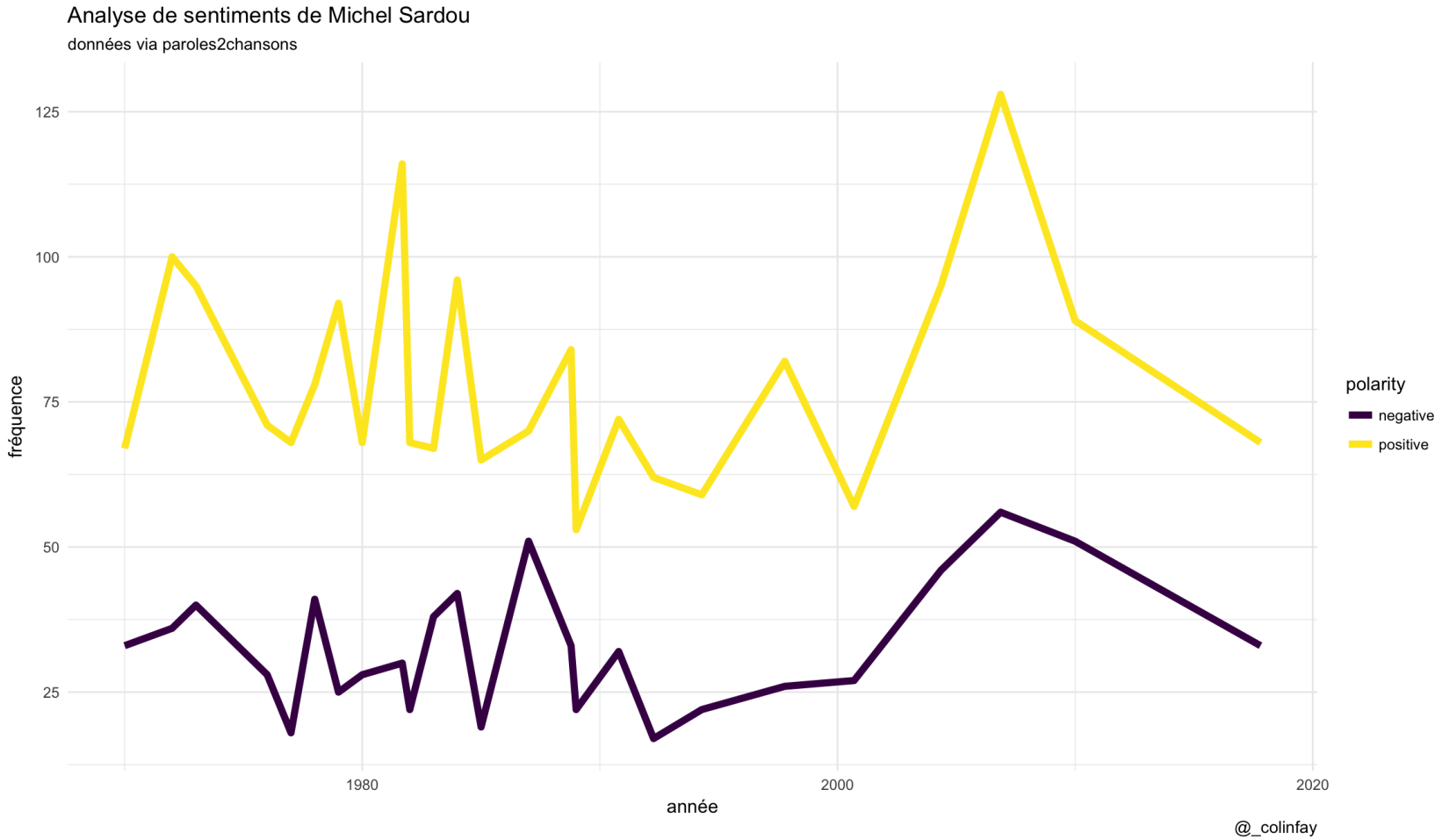
> C'est bien beau tout ça, et après ?

Après, on peut s'intéresser à d'autres choses dans la discographie de notre cher Michel.

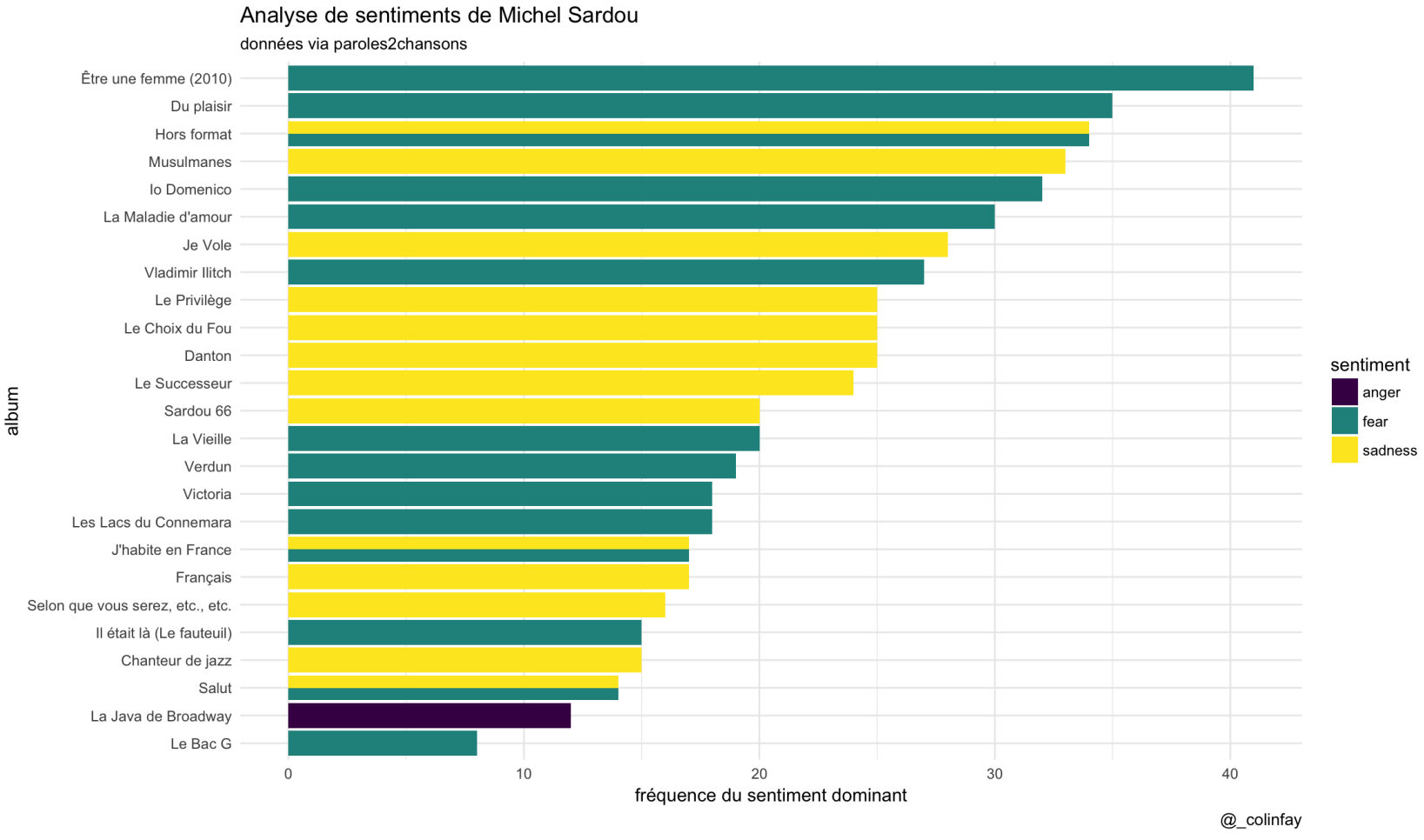
Par exemple, quels sont les albums les plus tristes ?

Et puisqu'on est ici, on va se faire un petit modèle de suggestion de chansons, tu en dis quoi ?

À travers les âges



Sentiment analysis



Topic Modeling

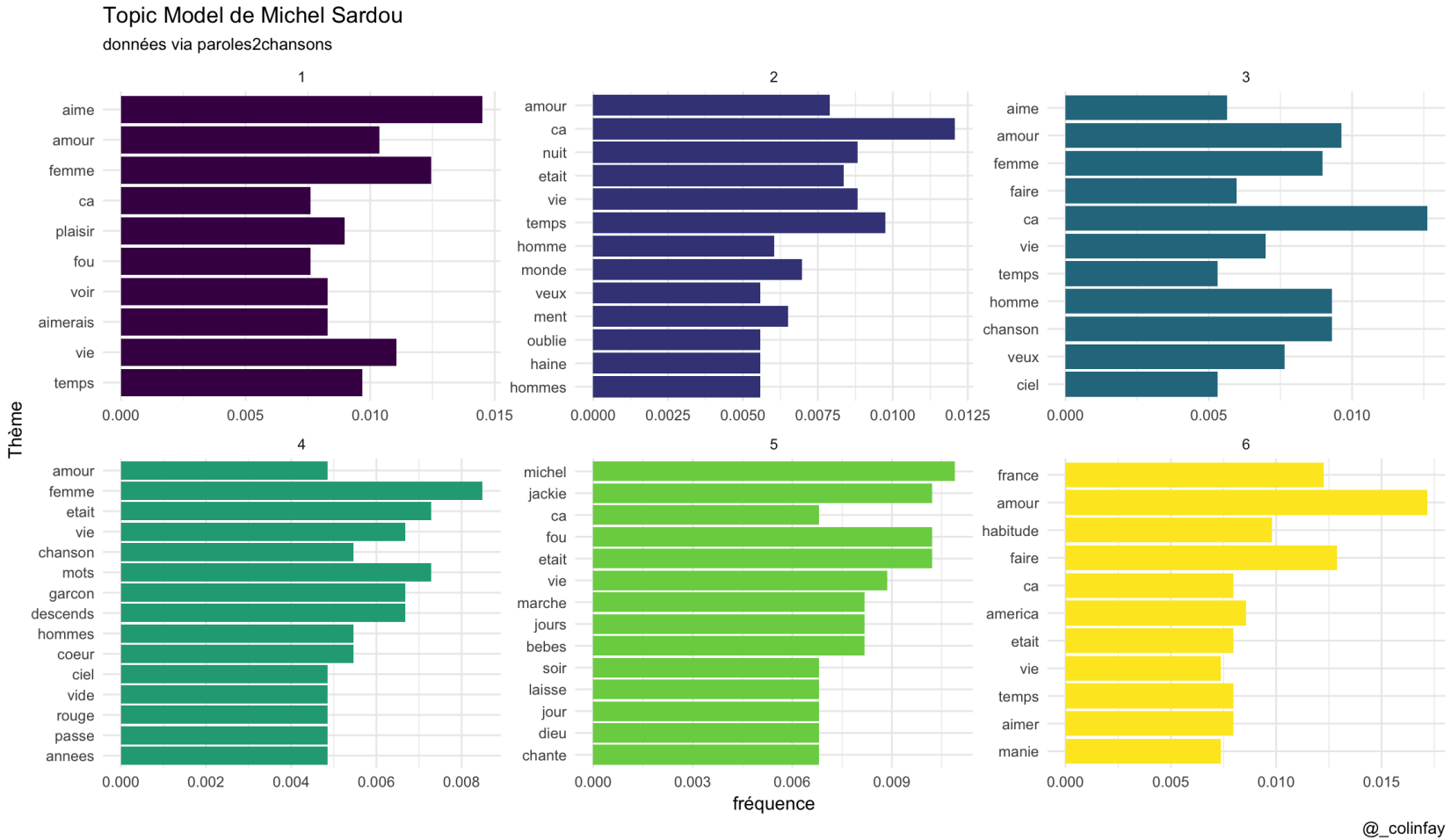
Bon, pour terminer, on va se faire un petit modèle statistique, qu'on appelle le "topic modeling".

Au delà d'avoir l'air ultra cool quand on dit ça, c'est un modèle qui permet de grouper des termes les uns avec les autres, afin d'en faire sortir des groupes.

Ensuite, on croise l'appartenance de chaque mot avec la chanson d'où il est tiré, ou de l'album.

Çe nous permet de faire des "clusters" de morceaux ou d'album, pour pouvoir te suggérer une chanson à partir d'une autre.

Topic models



Topic models

> Pas très clair tout ça

Ici, on a regroupé les termes qui participent le plus à la création de chaque groupe.

En clair, le premier topic a une forte présence de la question de la folie, de la vie, du temps.

Dans le second, on retrouve plutôt "haine", "oublie", "ment".

Et ainsi de suite.

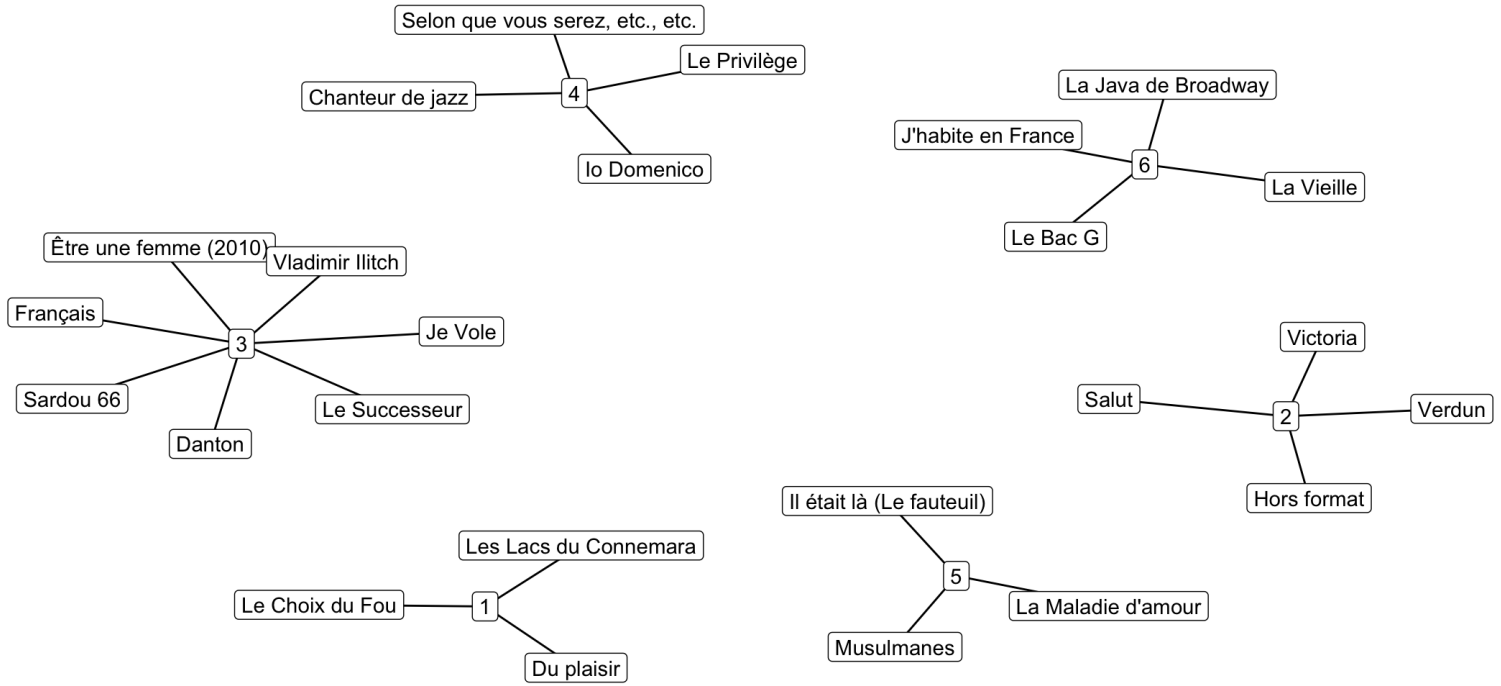
> Bon, maintenant, il faut qu'on puisse croiser ces infos de topic avec les albums !

Bien vu l'aveugle.

Topic models

Groupement des albums de Michel Sardou

données via paroles2chansons



@_colinfay

Suggestions

```
conseil_moi("Nuits Blanches A Rio")
```

chanson	album	date
Maman	Il était là (Le fauteuil)	1982-01-01
America America	J'habite en France	1970-01-01
Interdit aux bébés	La Maladie d'amour	1973-01-01
Corsica	Français	2000-09-12
J'accuse	La Vieille	1976-01-01
Danton	Danton	1972-01-01
Les dimanches	J'habite en France	1970-01-01
Le chanteur des rues	Le Bac G	1992-04-05
X Ray	Verdun	1979-01-01
8 jours à El paso	Je Vole	1978-01-01

Suggestions

```
conseil_moi("Etre une femme")
```

chanson	album	date
Le rire du sergent	J'habite en France	1970-01-01
Rouge	Io Domenico	1984-01-01
Qui m'aime me tue	Le Choix du Fou	2017-10-20
Une Femme Extraordinaire	Être une femme (2010)	2010-01-01
Marie-Jeanne	Le Privilège	1990-10-17
Attention les enfants... danger	Le Successeur	1988-10-15
Il était là	Il était là (Le fauteuil)	1982-01-01
Le grand réveil	Le Bac G	1992-04-05
Les mots d'amour	Chanteur de jazz	1985-01-01
La bataille	Français	2000-09-12

Suggestions

```
je_me_sens("joy")
```

chanson	album	date
Cette chanson-la	Français	2000-09-12
Il était là	Il était là (Le fauteuil)	1982-01-01
Merci... pour tout	Il était là (Le fauteuil)	1982-01-01
En chantant	Je Vole	1978-01-01
C'est ma vie	La Java de Broadway	1977-01-01
Les vieux mariés	La Maladie d'amour	1973-01-01
La chanson d'Eddy	Le Bac G	1992-04-05
Etre une femme	Les Lacs du Connemara	1981-09-07
Je m'en souviendrai sûr'ment	Salut	1997-10-12
Les yeux d'un animal	Vladimir Ilitch	1983-01-01

Suggestions

```
je_me_sens("fear")
```

chanson	album	date
Ce n'est qu'un jeu	Du plaisir	2004-05-10
Elle Vit Toute Seule	Être une femme (2010)	2010-01-01
Et Puis Après	Être une femme (2010)	2010-01-01
Etre une Femme 2010	Être une femme (2010)	2010-01-01
L'humaine Différence	Être une femme (2010)	2010-01-01
Atmospheres	Io Domenico	1984-01-01
Et alors !	Le Choix du Fou	2017-10-20
Le privilège	Le Privilège	1990-10-17
Le mauvais homme	Les Lacs du Connemara	1981-09-07
Dossier D	Victoria	1980-01-01

Tout a été fait avec R

R est un logiciel de développement scientifique spécialisé dans le calcul et l'analyse statistique.

Il a été utilisé de A à Z :

- Web scrapping
- Text-mining
- Graphiques
- Modélisation
- Création des slides

Tout a été fait avec R

Les packages utilisés ici

- Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org
- Colin FAY (2017). proustr: Tools for Natural Language Processing in French. R package version 0.2.1. https://github.com/ColinFay/proustr
- Simon Garnier (2017). viridis: Default Color Maps from 'matplotlib'. R package version 0.4.0. https://CRAN.R-project.org/package=viridis
- Thomas Lin Pedersen (2017). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. R package version 1.0.0. https://CRAN.R-project.org/package=ggraph
- Silge J and Robinson D (2016). "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS*, *1*(3). doi: 10.21105/joss.00037 (URL: http://doi.org/10.21105/joss.00037), .
- Hadley Wickham (2017). tidyverse: Easily Install and Load 'Tidyverse' Packages. R package version 1.1.1. https://CRAN.R-project.org/package=tidyverse

Thanks !

On garde les questions pour la fin ?

Find me on the web:

- colin@thinkr.fr
- http://twitter.com/_colinfay
- http://twitter.com/thinkr_fr
- https://github.com/ColinFay

And also:

- https://thinkr.fr/
- http://colinfay.me/