Université Paul Valéry - Montpellier 3

UFR 6 - Éducation & Sciences pour les LLASHS

Master MIASHS

Prédiction des comportements suicidaires

RENDU 8 - VISUALISATION

Équipe :
Lisa BÉTEILLE
Matéo CALSACY
Jean CHABANOL
Anamé ROUMY
Laura SÉNÉCAILLE
Célia TEYSSIER

2021/2022

Encadrants :
Sandra BRINGAY
Pierre LAFAYE de MICHEAUX
Florian LOMBARDO

SOMMAIRE

TABLE DES FIGURES

INTRODUCTION

La visualisation permet de renforcer un message en insistant sur les éléments visuels. Elle nous permet de nous familiariser avec les données, nous oriente dans nos recherches et rend compréhensible ces données. Le rôle de la visualisation nous permet d'éclairer les informations récoltées en les traduisant par des représentations visuelles claires et accessibles, tout en étant plus puissant et persuasif qu'un rapport.

"Une image vaut mille mots" de Confucius représente parfaitement l'intérêt de la data visualisation.

Le but de la visualisation au travers de notre sujet sera de connaître le vocabulaire ou les familles de mots les plus utilisés autour du thème du suicide ou de la vie, tout en nous familiarisant avec ces données pour nous rendre efficace durant ce projet.

Au premier semestre, nous avons commencé la visualisation sur les données avec des nuages de mots et des analyses de similitudes, avant d'aller vers des visualisations au second semestre au travers de visualisations plus pertinentes grâce au modèle de BERT.

PARTIE 1 - VISUALISATION DU 1ER SEMESTRE

Pour réaliser les visualisations du semestre 1, nous avons utilisé Iramuteq. Les visualisations qui nous ont paru les plus pertinentes à faire avec des corpus de textes ont été des nuages de mots et des analyses de similitude. Cela nous a permis d'extraire du vocabulaire important souvent présent dans la plupart de nos données pour nos analyses. Cela nous a aussi permis d'explorer les données pour se familiariser avec et connaître les sujets principaux concernant les tentatives de suicide sur les plateformes telles que Reddit, Wikipédia et Twitter.

I. Graphique de prétraitement des données

a. Wikipédia

Dans les nuages de mots ci-dessous, on remarque une grande différence de vocabulaire entre les deux catégories de textes. Dans les textes biographies de personnes suicidées (figure 2), les notions de mort et de suicide sont très présentes ainsi que celles de la famille. Dans les textes des personnes n'ayant pas fait de tentative de suicide (figure 1), on retrouve de façon plus inattendue une grande présence d'adjectifs tels que young, early, old, short.

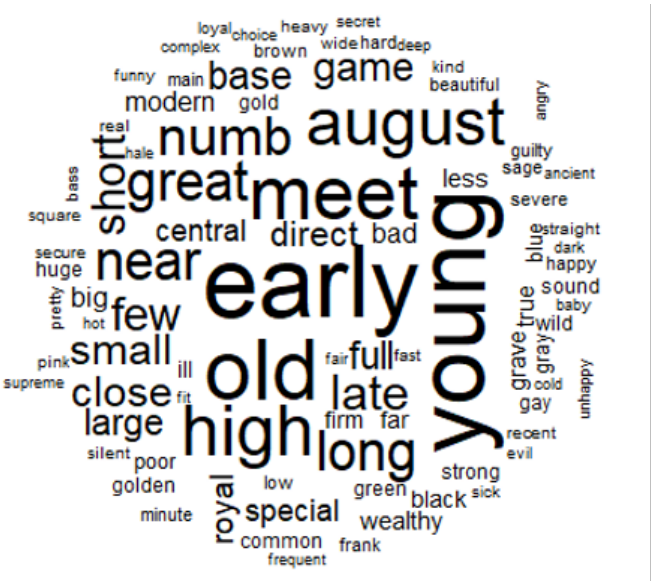

Figure 1 : Nuage de mots formes actives - Wikipédia - List of actors, List of actress, List of artistes

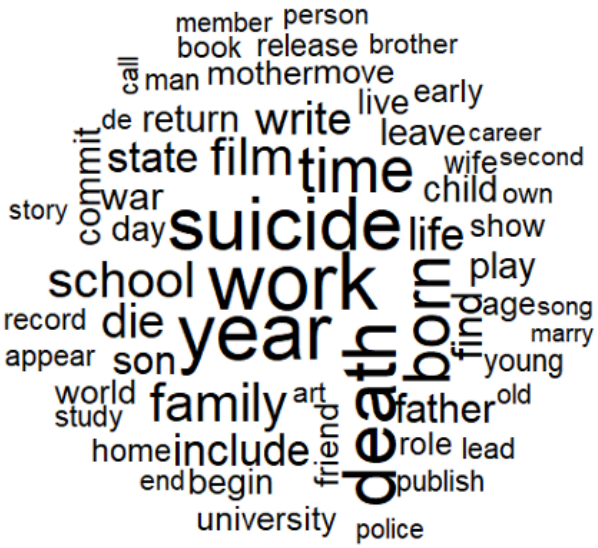

Figure 2 : Nuage de mots formes actives - Wikipédia - List of suicides

L'analyse des similitudes sur les biographies Wikipédia des personnes qui se sont suicidées montre plusieurs familles de mots (figure 3) : une notion de travail, une notion de famille, une notion de parcours scolaire qui sont assez attendues pour des biographies mais aussi des notions de suicides et de guerres.

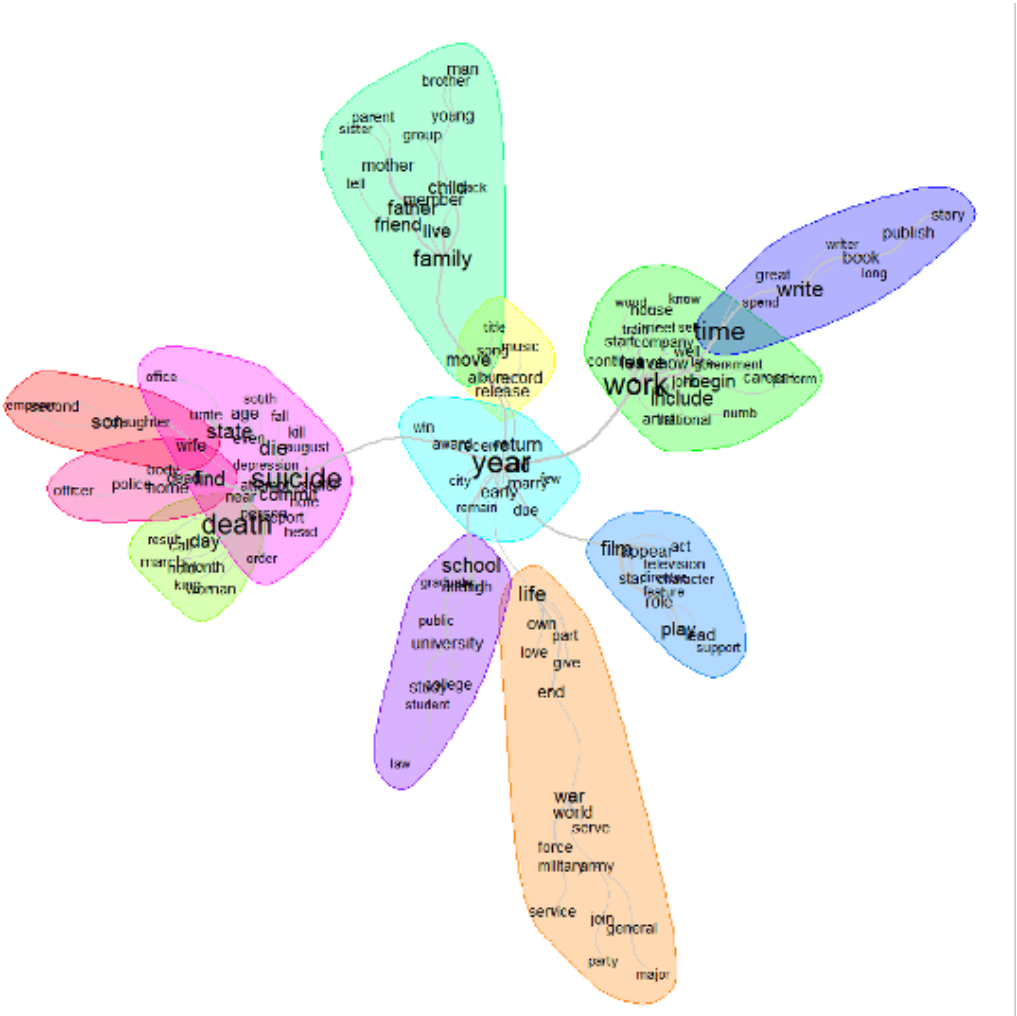

Figure 3 : Analyse des similitudes - Wikipédia - List of suicides

b. Reddit

Dans les données provenant de Reddit, on remarque une similitude dans les vocabulaires des deux catégories : les messages sont centrés sur les ressentis personnels et sur les émotions (figure 4 et 5).

Figure 4 : Nuage de mots formes actives - Reddit – TS

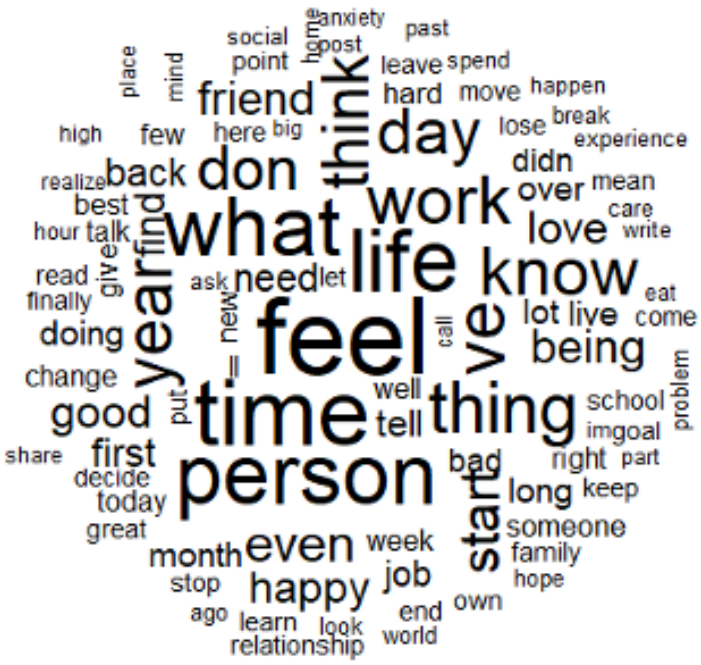

Figure 5 : Nuage de mot formes actives - Reddit – non TS

D’ailleurs, si on se concentre sur les verbes les plus utilisés dans les deux catégories (figure 6 et 7) de Reddit on trouve surtout feel (ressentir) ou know (savoir) et pratiquement les mêmes verbes pour les tentatives de suicides et les non tentatives de suicides.

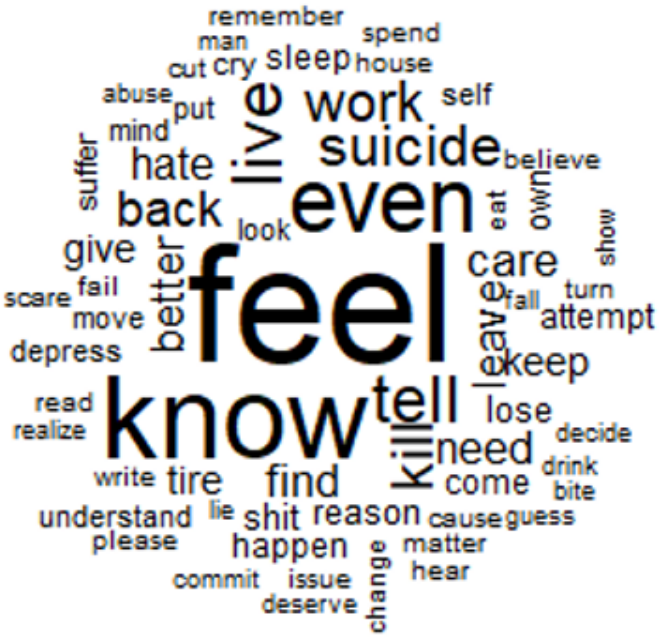

Figure 6 : Nuage de mots verbes - Reddit – TS

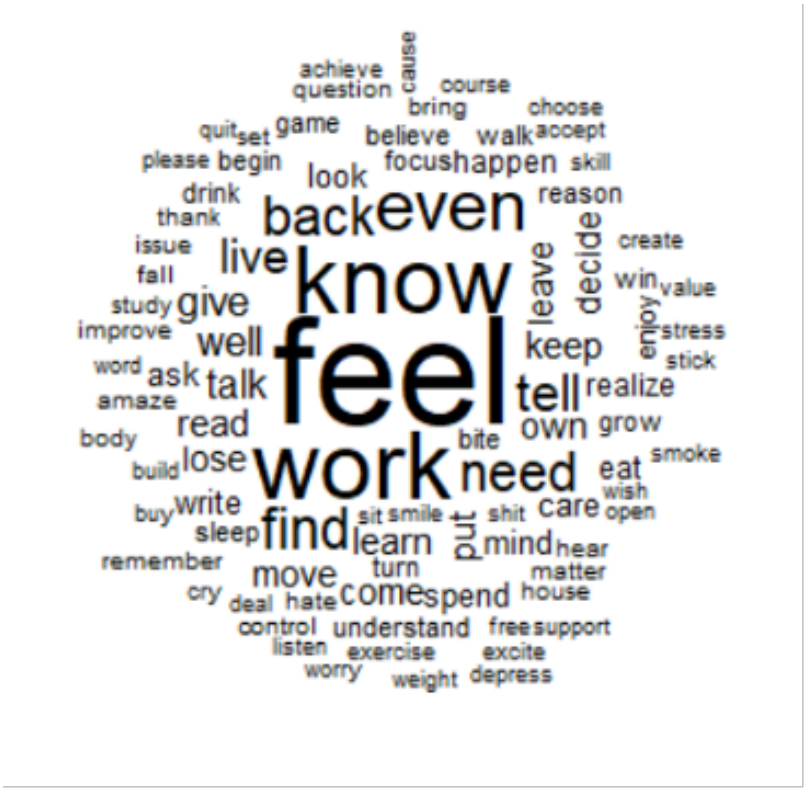

Figure 7 : Nuage de mots verbes - Reddit non TS

De façon peu surprenante, suite aux nuages de mots précédents, on trouve le mot me (moi) en racine principale dans l'analyse des similitudes des textes liés à des tentatives de suicides de Reddit (figure 8).

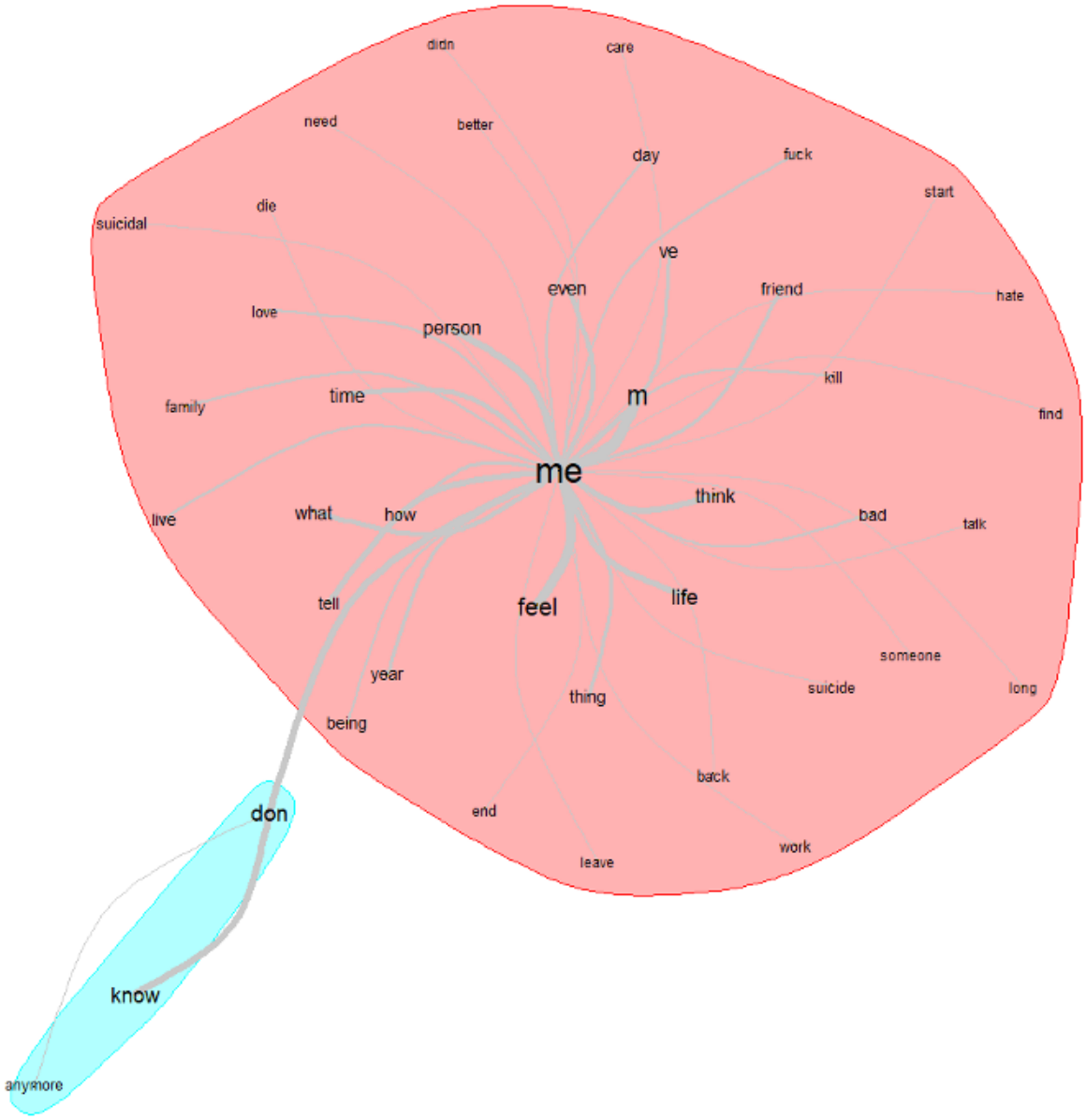

Figure 8 : Analyse des similitudes - Reddit - TS

c. Twitter

Dans les textes de Twitter, les textes semblent très polarisés. Dans la catégorie suicide (figure 9), les mots bully, suicide, die et kill sont assez marquants. En comparaison, on trouve beaucoup de vocabulaire joyeux ou plus général dans la catégorie non tentatives de suicides comme happy, news, day, … (figure 10).

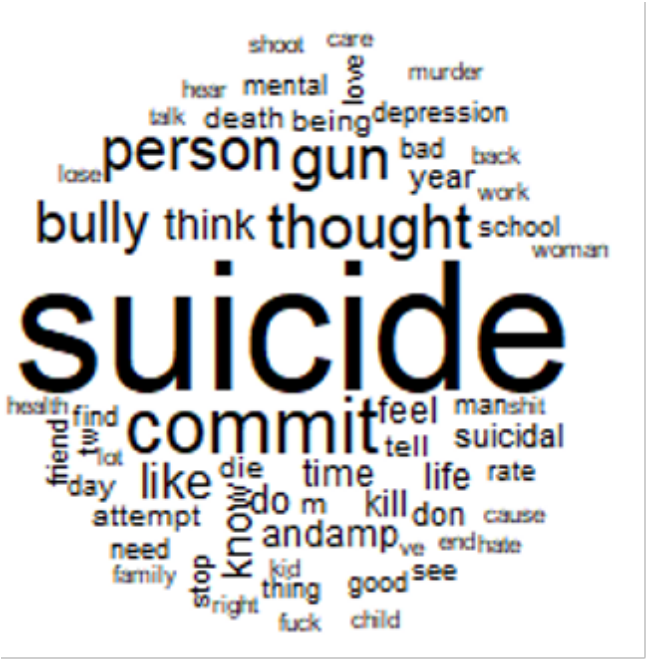

Figure 9 : Nuage de mots formes actives - Twitter – TS

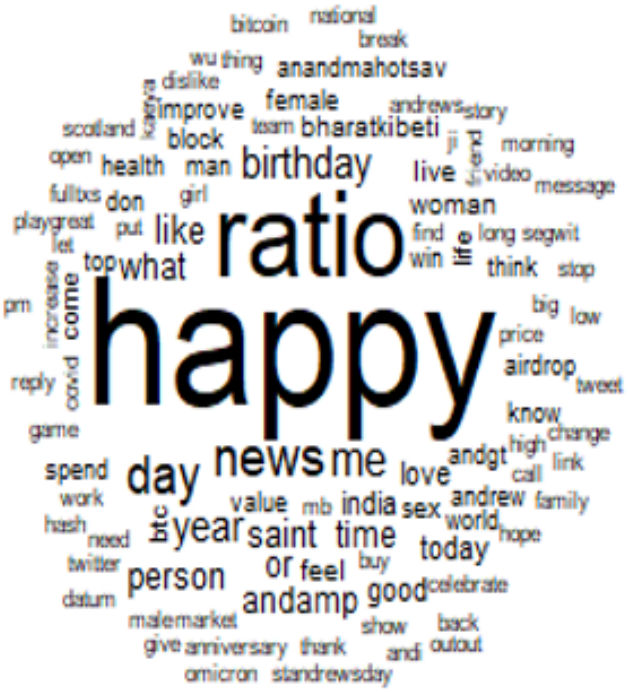

Figure 10 : Nuage de mots formes actives - Twitter – non TS

L’analyse des similitudes des publications de Twitter liées au suicide (figure 11) montre une famille de mots principale liée au suicide et au mal-être, mais on voit aussi apparaitre des notions de harcèlement scolaire, de meurtre et de maladie mentale qui ne sont pas inattendues.

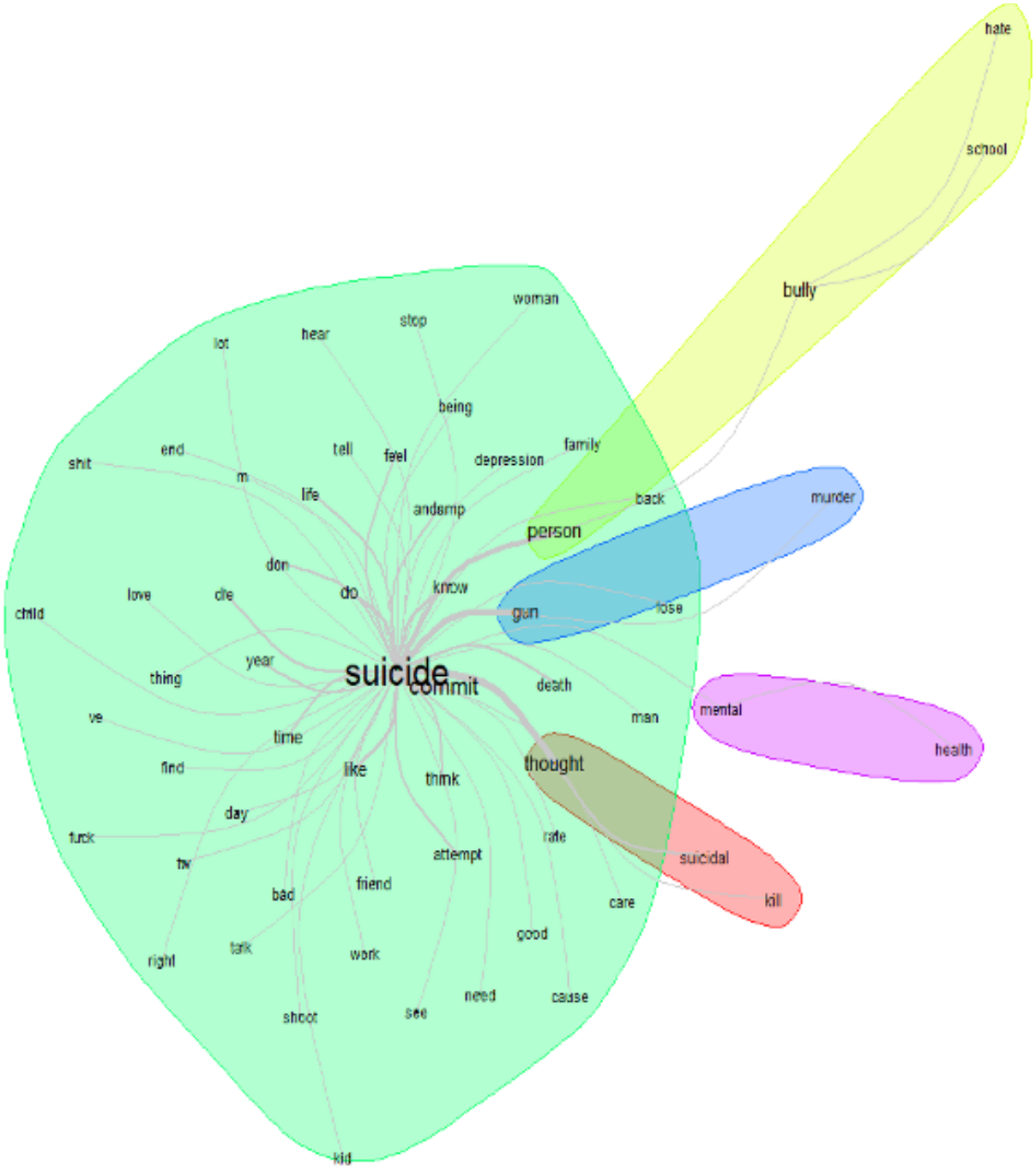

Figure 11 *: Analyse des similitudes - Twitter - TS*

PARTIE 2 - VISUALISATION DU 2ÈME SEMESTRE

Explorer un corpus c'est aussi explorer ses sujets prédominants, analyser la similarité entre les textes et voir comment ils se regroupent. En effet, au travers de notre modèle nous cherchons à prédire si la personne rédigeant un texte est suicidaire ou non en s'appuyant sur des données Wikipédia. Il nous semble donc pertinent d'analyser les éventuels groupes de sujets que pourraient former nos données Wikipédia.

BERTopic est une technique de modélisation de sujets qui exploite les intégrations BERT et c-TF-IDF pour créer des clusters denses permettant des sujets facilement interprétables tout en conservant les mots importants dans les descriptions de sujets.

TF-IDF permet de comparer l'importance des mots entre les textes en calculant la fréquence d'un mot dans un texte donné et aussi la mesure de la prévalence du mot dans l'ensemble du corpus.

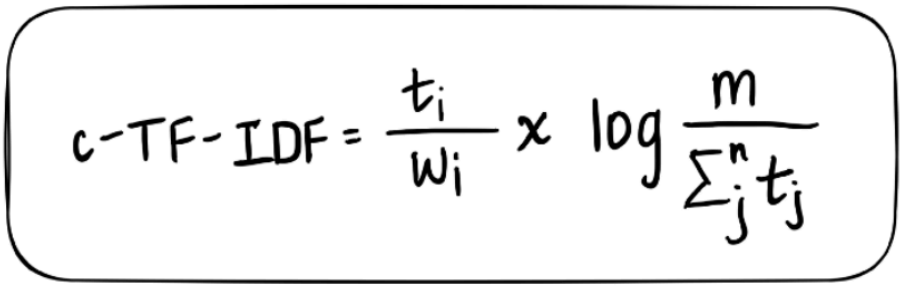

Image made by author using excalidraw | c-TF-IDF formula: the la fréquence de chaque mot t est extraite pour chaque classe i et divisée par le nombre total de mots w. Le nombre moyen de mots par classe m est divisé par la fréquence totale des mots t dans toutes les n classes.

Maintenant, si nous traitons plutôt tous les textes d'un seul cluster comme un seul texte, puis effectuons TF-IDF, le résultat serait des scores d'importance pour les mots d'un cluster. Plus les mots sont importants dans un groupe, plus ils sont représentatifs de ce sujet. Par conséquent, nous pouvons obtenir des descriptions basées sur des mots-clés pour chaque sujet.

La carte de distance intertopique est une visualisation des sujets dans un espace à deux dimensions. La superficie de ces cercles thématiques est proportionnelle au nombre de mots appartenant à chaque thème dans le dictionnaire. Les cercles sont tracés à l'aide d'un algorithme de mise à l'échelle multidimensionnelle (convertit un tas de dimensions, plus que ce que nous pouvons concevoir avec notre cerveau humain, en un nombre raisonnable de dimensions, comme deux) en fonction des mots qu'ils comprennent, de sorte que les sujets qui sont plus proches les uns des autres ont plus de mots en commun.
Nous avons choisi de déterminer vingt sujets, représentés ici par des cercles, qui se sont rassemblés en 6 groupes. Chaque groupe de cercle est un groupe de sujet qui traduit la proximité de ces derniers par le fait qu'ils aient beaucoup de mots en commun.
La barre en bas du graphique nous permet de repérer un topic dans le graphique en le colorant en rouge.

***Figure 12** : représentation des sujets de nos corpus regroupés selon leurs similitudes*

La carte de distance intertopique ci-dessus représente 6 principaux groupes de sujets parmi nos 20 sujets au total.

Bien que l'on ait montré comment accéder aux principaux mots-clés appartenant à un sujet particulier et à leurs scores d'importance, nous pouvons également visualiser ces termes et scores sous forme de graphiques à barres. Voici les 8 premiers topics :

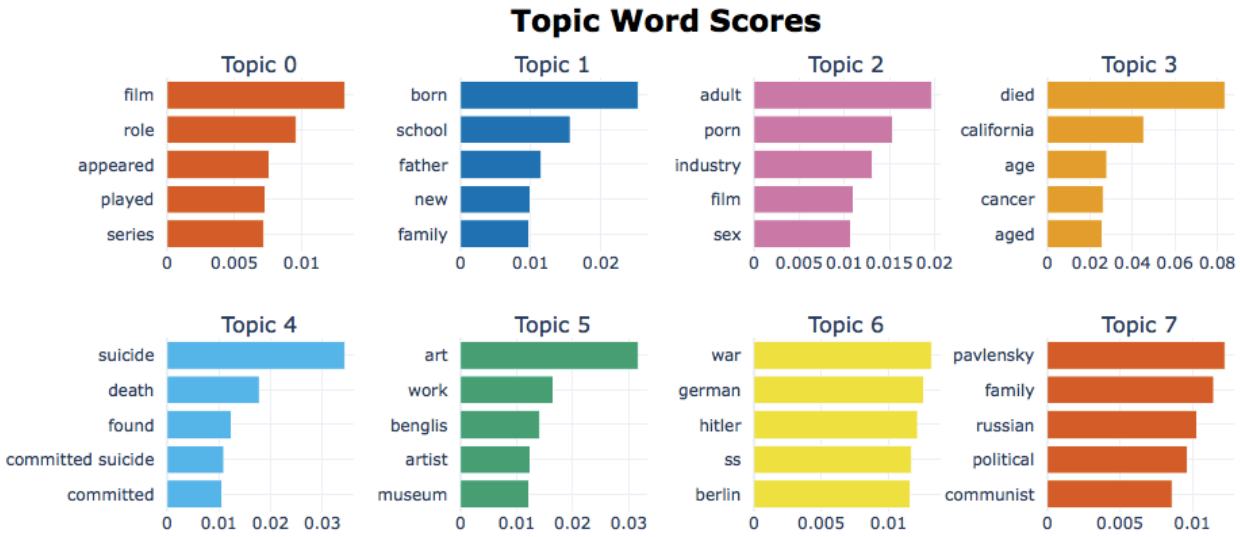

***Figure 13** : représentation de 8 de nos 20 sujets avec les 5 mots les représentant le mieux*

Le graphique à barres affiche par défaut les 5 termes les plus saillants. Les barres indiquent la fréquence totale du terme sur l'ensemble du corpus. Saillant est une métrique spécifique, définie au bas de la visualisation, qui peut être considérée comme une métrique utilisée pour identifier les mots les plus informatifs ou utiles pour identifier les sujets dans l'ensemble de la collection de textes. Des valeurs de saillance plus élevées indiquent qu'un mot est plus utile pour identifier un sujet spécifique.

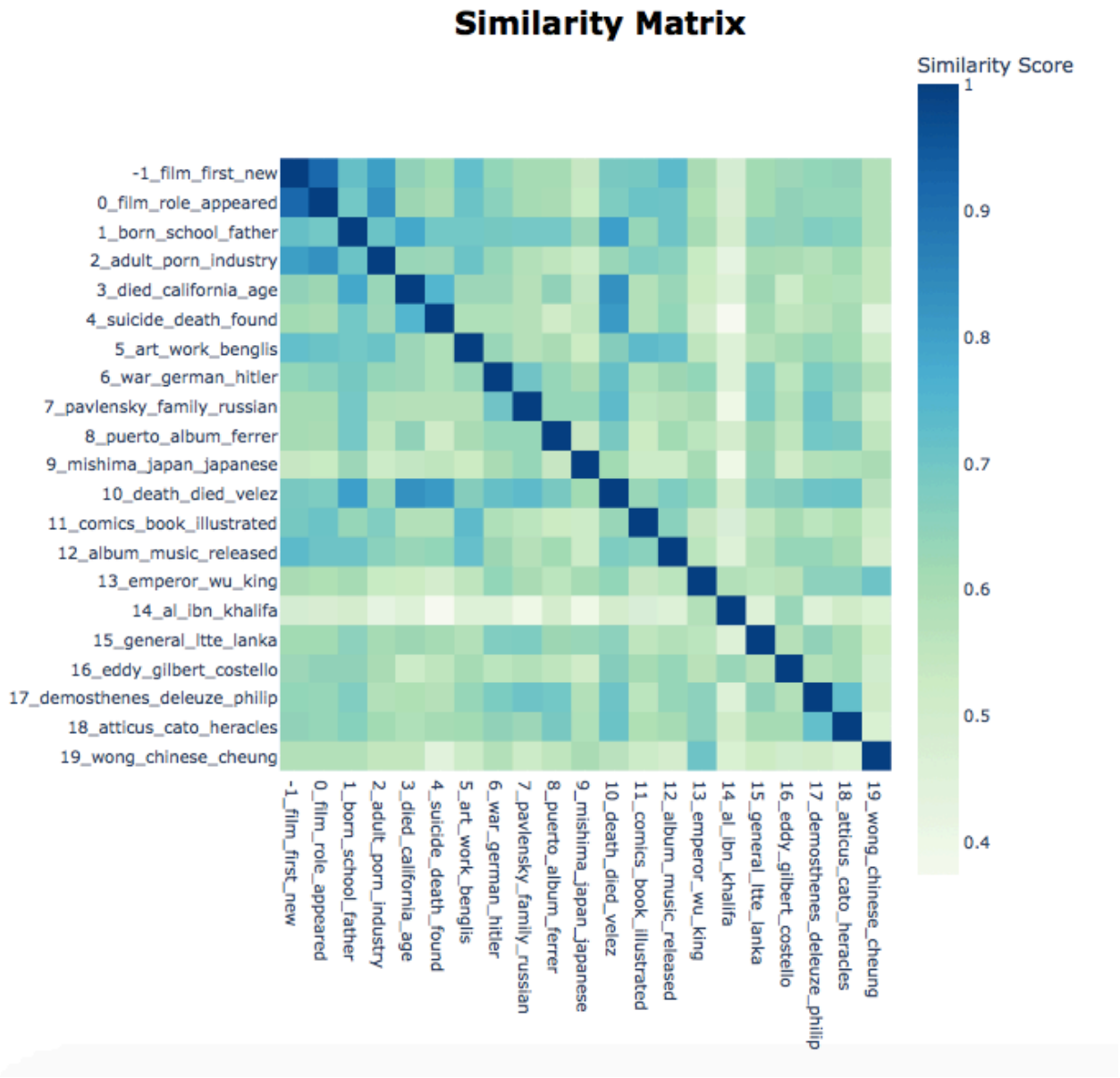

Figure 14 *: Matrice de corrélation entre nos différents sujets*

Ici la matrice de corrélation entre les sujets. Sur chacun des axes nous avons nos vingt sujets. Plus c'est proche de 1 plus les sujets sont proches, ce qui crée une diagonale de 1 quand on calcule la corrélation entre deux mêmes sujets. Sur la matrice, plus la case est foncée, plus les deux sujets sont corrélés. Par exemple, les sujets 3 et 4 sont assez corrélés puisque la case entourée en rouge est bien bleue. Ceci est logique puisque sur la figure 1, ces deux sujets font partis du même groupe, celui en bas à gauche. On devine donc alors que les sujets avec une forte corrélation feront partis de même cluster ! Il est possible de zoomer sur la matrice.

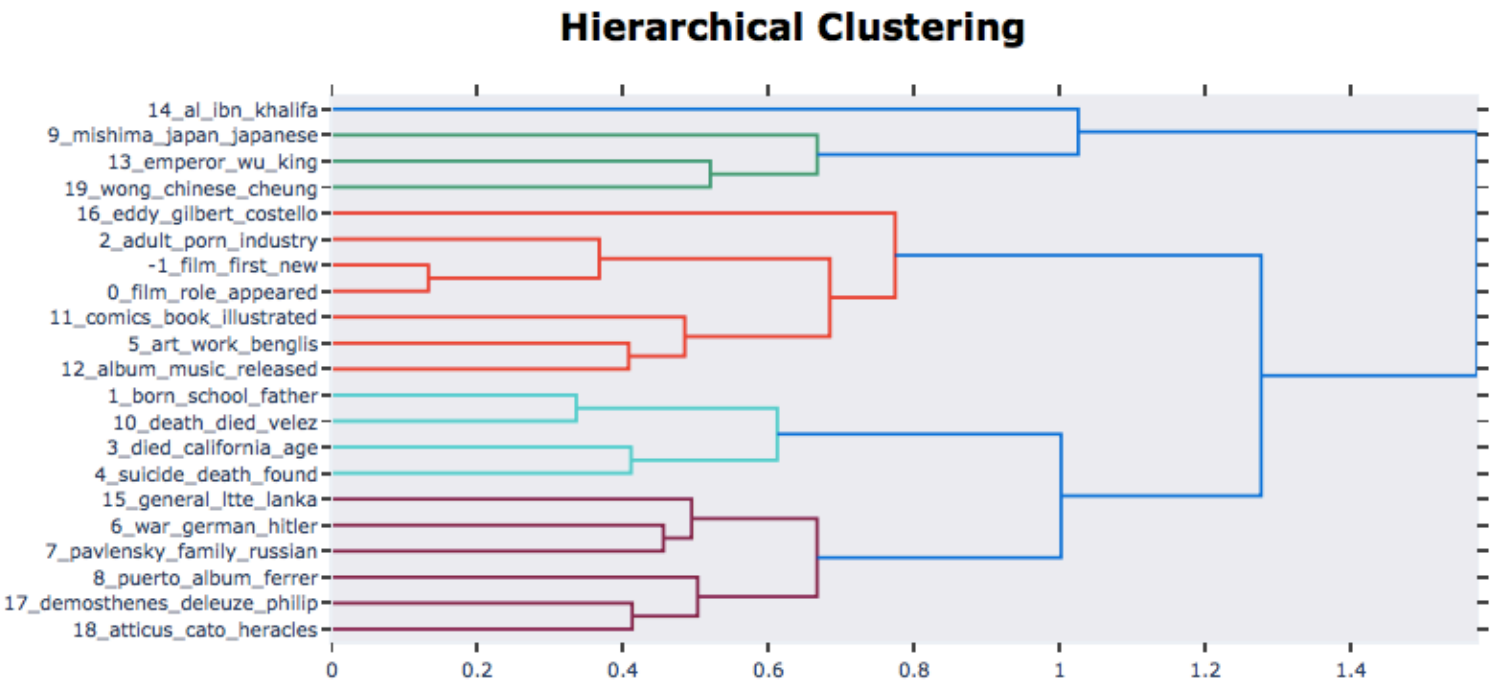

***Figure 15** : représentation des sujets de nos corpus regroupés selon leurs similitudes*

Ce graphique montre la même chose que celui de la figure 12. Un dendrogramme est un type de diagramme en arbre montrant les relations hiérarchiques entre différents ensembles de données.Voici la hiérarchisation des groupes de sujets. La distance entre les points de données représente les dissemblances. La hauteur des blocs représente la distance entre les clusters.

Nous avons zoomé pour voir plus en détail mais cette visualisation nous montre nos 6 groupes principaux de sujets tous d'une couleur différente. Ici, nous observons qu'il y a 5 groupes et non 6, car sur le graphique intertopic distance map (figure 12), le dendrogramme a réuni deux groupes en un seul. Ce groupe correspond aux branches rouges de la figure ci-dessus et aux deux groupes de la figure 12 représentés ci-dessous.

***Figure 16** : Deux groupes de la figure 1, correspondant à un sujet globale : L'art*

Les cinq groupes de nos vingt sujets correspondent aux sujets globaux :

- La série Khalifa (Bleu)
- La Chine et le Japon (Vert)
- Les arts, films, musiques, livres… (Rouge)
- La mort, le suicide (Bleu clair)
- La seconde Guerre mondiale, avec Hitler, les russes, Heracles (Opération Herkules qui est le nom de code d'un plan militaire pendant la seconde guerre mondiale) et Démosthène (dont les français se sont identifiés lors de la résistance et ont donné à Adolf Hitler le nom de Philippe). Le sujet general_Itte_lanka de ce groupe fait référence à la guerre civile au Sri Lanka et est donc associé à ce groupe puisqu'il parle aussi de guerre. (Rouge bordeaux)

Pour conclure simplement en quelques mots, nous pouvons être rassurés quant à l'exactitude de notre analyse sur les sujets car chacun de nos graphiques créés à partir de la librairie BERTTopic nous montrent en globalité la même chose. Nos données Wikipédia recensent environ 20 sujets différents qui peuvent se rejoindre de par leur similarité pour enfin créer 6 groupes de sujets.

CONCLUSION

Pour conclure, nous avons pu durant cette année explorer de nombreuses visualisations différentes qui ont permis de donner du sens à nos données. Nous avons également vu que ces visualisations se sont précisées au cours de l'année en commençant tout d'abord par des nuages de mots et des analyses de similitudes pour au final, nous permettre de découvrir de nouvelles technologies (BERTopic) en faisant la carte de distance intertopique ou les graphiques à barres par exemple.