Une image contenant texte

Description générée automatiquementC

Université Paul Valéry – Montpellier 3

UFR 6 – Education & Sciences pour les LLASHS

Master MIASHS

Prédiction des comportements

suicidaires

**RENDU 3 – DONNEES ET EXPLORATION DES DONNEES**



Equipe :

Lisa BETEILLE

Matéo CALSACY

Jean CHABANOL Encadrants :

Anamé ROUMY Sandra BRINGAY

Laura SENECAILLE 2021/2022 Pierre LAFAYE de MICHEAUX

Célia TEYSSIER Florian LOMBARDO

Avoir un jeu de données pertinentes est essentiel pour mener un projet de recherches à bien. Des milliers de données sont disponibles sur internet, encore faut-il pouvoir y accéder et faire le tri entre celles qui nous seront utiles et celles qui ne le seront pas. De plus, il savoir les exploiter.

Notre projet consiste à produire un ou plusieurs algorithmes qui permettront de déterminer si un texte peut avoir été écrit par une personne risquant de faire une tentative de suicide. Dans ce projet déjà commencé pendant les années précédentes par les promotions de MIASHS, les jeux de données sur lesquels nous travaillons étaient déjà déterminés : des messages publics venant des réseaux sociaux Reddit et Twitter, et des biographies de Wikipédia.

Ce rapport est consacré aux données utilisées dans ce projet et à leur exploration.

Comme expliqué dans l’introduction, nos sources de données étaient prédéterminées :

* des biographies en anglais venant de Wikipédia,
* des publications venant des fils de conversations de Reddit,
* des «tweets » provenant de Twitter.

Ces trois sources de données sont assez différentes. Wikipédia fournit des biographies de personnes célèbres dans un langage assez soutenu et codifié, on va notamment retrouver des écrivains et des acteurs. L’intérêt de ces biographies serait de pouvoir identifier les personnes qui se sont suicidées ou qui pourraient faire des tentatives en fonction de leur parcours de vie. Ces données sont très faciles d’accès, les commanditaires du projet nous ont fourni un notebook python qui permet d’extraire toutes les biographies d’une catégorie choisie de Wikipédia. De plus, il est aisé de trouver les personnes qui se sont suicidées, elles sont listées dans une catégorie du site. Nous avons néanmoins eu quelques difficultés suite à des mises à jour de Wikipédia qui ont temporairement empêché l’algorithme de fonctionner.

D’un point de vue moral et éthique, utiliser des biographies Wikipédia est tout à fait acceptable : ce sont des données publiques accessibles à tous. Autres avantages de cette source de données : le coût de l’extraction est nul, nous pouvons extraire des pages à l’infini, dans la limite des biographies disponibles, et de façon assez rapide, environ une quinzaine de minutes.

Dernier avantage : les données Wikipédia sont extraite directement dans un tableau au format .csv aisé à manipuler à l’aide de R ou de python. L’extraction fournit les variables suivantes :

|  |  |  |  |
| --- | --- | --- | --- |
| Dictionnaire des variables | | | |
| Nom | Code | Contenu | Format |
| Lien de la page | pages\_links | Lien de la page wikipédia | string |
| Nom de la page | pages\_names | Nom de la personne | string |
| Sous-catégorie | subcategory | Sous catégorie wikipédia | string |
| Contenu | content | Biographie de la personnne | string |

Notre deuxième source de données est Reddit. Les données proviennent textes publiés sur des fils appelés des « subreddit ». Chaque fil possède un thème, ce qui est très pratique pour identifier les publications à caractère suicidaire ou non en fonction de ce fil, même si nous ne pouvons pas avoir une certitude exacte que toutes les publications sont bien écrites par des personnes suicidaires. Ces données nous sont fournies directement par nos commanditaires. Cette source de données possède une limite : dans le cas d’une extraction de données en lien avec le suicide, seules les publications des fils en lien avec des tentatives de suicides sont extraites. Il n’est donc pas possible de demander plus de données à notre commanditaire. C’est néanmoins notre groupe qui a choisi les fils dont les publications sont extraites. Ces données ont un grand intérêt : elles sont écrites directement par des personnes qu’on suppose susceptibles de faire des tentatives de suicide. Le format d’extraction des publications n’est pas le plus pratique pour être immédiatement manipulé : ce sont en effet des extractions au format .json. L’extraction fournie notamment les variables suivantes :

|  |  |  |  |
| --- | --- | --- | --- |
| Dictionnaire des variables | | | |
| Nom | Code | Contenu | Format |
| Identifiant | id | Identifiant de la publication | String |
| Publication | text | Contenu de la publication | String |
| Titre | title | Titre de la publication | String |
| Nom du fil | subreddit | Nom du fil d'origine de la publication | String |
| Date de création | creation | Date de création de la publication | Date |

Les individus s’exprimant sur Reddit le font de leur plein gré et ont connaissance du caractère public de leurs publications. Il est acceptable d’utiliser cette source de données pour notre projet. Néanmoins, au regard du charactère très personnel de certaines des publications et de la méconnaissance des usages possibles de leurs données, surtout de la part des plus jeunes, on peut se demander si les auteurs des différentes publications récoltées auraient donné leur accord pour cet usage. De plus, même si les publications sont théoriquement anonymes, certaines d’entre elles contiennent peut-être des éléments permettant de retrouver l’identité de son auteur, le jeu de données contient potentiellement des données personnelles même si c’est en nombre infiniment faible.

Le site Kaggle disposait également de certains jeux de données moins récents pour Reddit sur les mêmes fils. Le coût de cette extraction et le temps qu’elle a nécessité n’est pas connu par notre groupe.

Notre dernière source de données provient de Twitter. Ces tweets nous sont fournis par nos commanditaires. Les tweets liés à des tentatives de suicide sont ceux fournis l’année dernière au groupe précédent en charge de ce projet et sont extraits sur une courte période. Les tweets non liés à des tentatives de suicide ont été fourni cette année par nos commanditaires. Les données extraites sont toutes au format .json mais il faut remarquer que la structure interne des documents importée n’est pas la même en fonction du moment où elles sont étés importées. De façon synthétique, voici les différents noms de variables utiles et leur contenu :

|  |  |  |  |
| --- | --- | --- | --- |
| Dictionnaire des variables | | | |
| Nom | Code | Contenu | Format |
| Identifiant | id | Identifiant du tweet | String |
| Date de création | origin | Date de création du tweet (nouvelle version) | Date |
| Date de création | created\_at | Date de création du tweet (ancienne version) | Date |
| Contenu | text | Contenu du tweet (ancienne version) | String |
| Contenu | full\_text | Contenu du tweet (nouvelle version) | String |

Tout comme pour Reddit, les publications provenant de Twitter sont volontairement écrites de façon publique et expriment un point de vue personnel intéressant à utiliser dans la prédiction de comportements suicidaires. Pour départager les publications écrites par des personnes ayant supposément fais des tentatives de suicide et les autres, on considère que certains mots sont révélateurs d’une tentative de suicide tels que suicide, depressed, gun,… Cette façon de départager les publications est la seule que nous ayons même s’il est vrai qu’avoir écrit une publication contenant le mot suicide n’implique pas que nous allons faire une tentative de suicide.

Les données provenant de Twitter sont difficiles à extraire, puisque seuls les tweets provenant de moins d’une semaine peuvent être exportés. Demander plus de publications provenant de Twitter à nos commanditaires pourrait nécessiter l’utilisation de beaucoup de ressources de leur part et prendre un certain nombre de temps. Nous ne connaissons pas le coût de cette opération.

Les questions d’éthique et de confidentialité à se poser sont les mêmes que celles pour les données provenant de Reddit.

Pour entrainer nos algorithmes de prédiction, nous avons choisi d’utiliser de mettre en forme et de changer le format des jeux de données à l’aide du logiciel R. Les jeux de données ont tous été passés au format .csv pour faciliter leur import sur des notebooks python qui contiennent les algorithmes de prédiction. Ils ont aussi été modifiés pour correspondre au format de lecture du logiciel Iramuteq qui a servit à faire la description des données.

Le logiciel R nous a aussi permis de nettoyer les jeux de données, en enlevant notamment les caractères non lisibles, les apostrophes et d’autres caractères de ponctuation non nécessaires.

Le tableau suivant apporte une première description générale des jeux de données :



*TS = Tentative de Suicide*

*Nombre d’occurrence = Nombre de mots*

*Nombre de formes = Nombre de mots différents*

*Nombre d’hapax = nombre de mots trouvés seulement une fois*

*Nombre moyen d’occurrence par texte = Nombre moyen de mots par texte*

On remarque que généralement nous possédons moins de textes labellisés comme TS que de non TS, les jeux de données sont rééquilibrés avant d’être utilisés dans les algorithmes de prédiction. Fait remarquable : les textes provenant de Twitter sont beaucoup plus courts en moyenne que ceux de Reddit et de Wikipédia.

Analysons maintenant les jeux de données un par un. Nous commençons par les textes Wikipédia

Dans les nuages de mots ci-dessous, on remarque une grande différence de vocabulaire entre les deux catégories de textes. Dans les textes de personnes ayant fait des TS, 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 TS, on retrouve de façon plus inattendu une grande présence d’adjectifs tels que young, early, old, short,..

Une image contenant texte

Description générée automatiquementUne image contenant texte

Description générée automatiquement

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 : une notion de travail, une notion de famille, une notion de parcours scolaire qui sont assez attendues pour des biographie mais aussi des notions de suicides et de guerres.

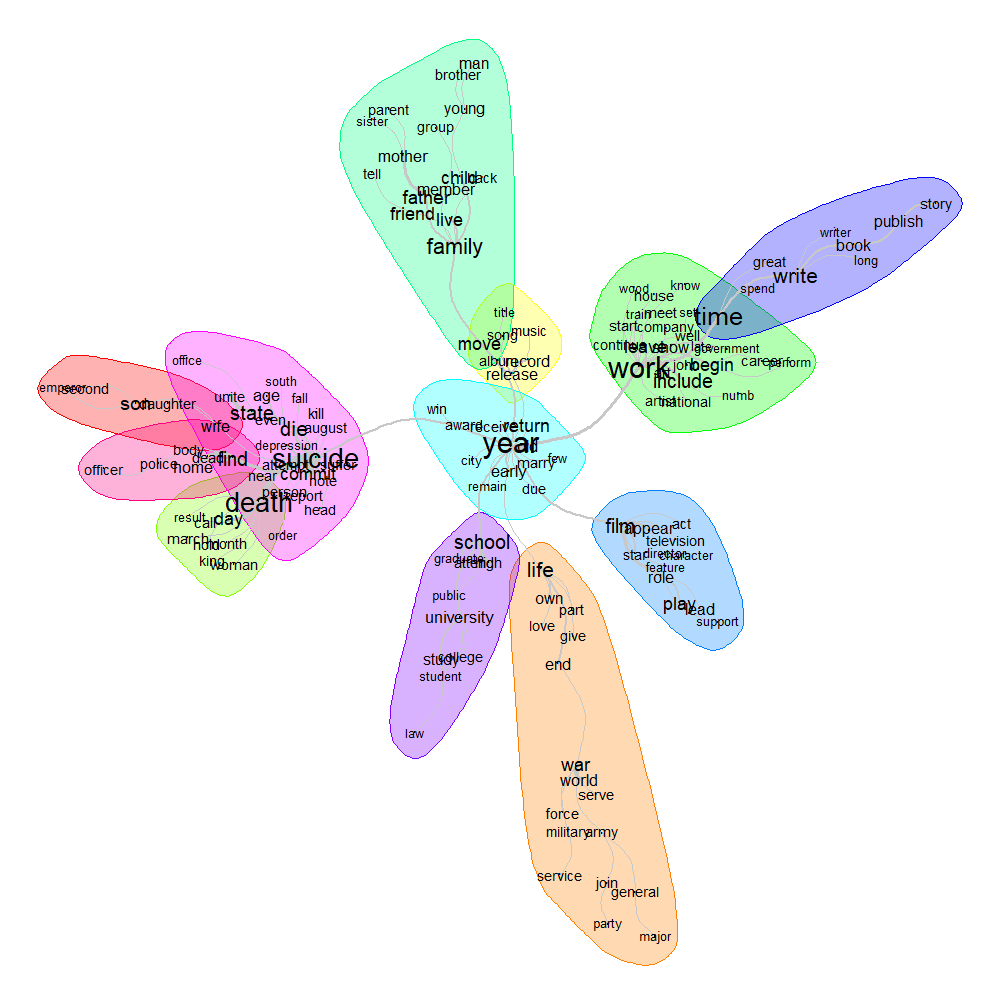


Figure 3 : Analyse des similitudes - Wikipédia - List of suicides

Dans les textes de Twitter, les textes semblent très polarisés. Dans la catégorie suicide, les mot bully, suicide, die et kill sont assez marquant. En comparaison, on trouve beaucoup de vocabulaire joyeux ou plus général dans la catégorie non TS comme happy, news, day,…

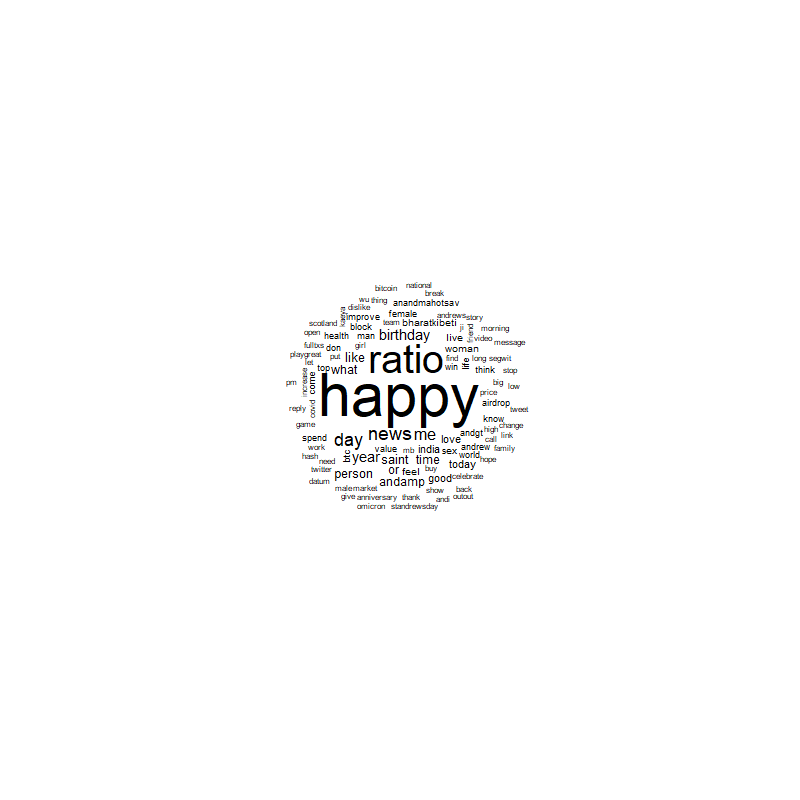
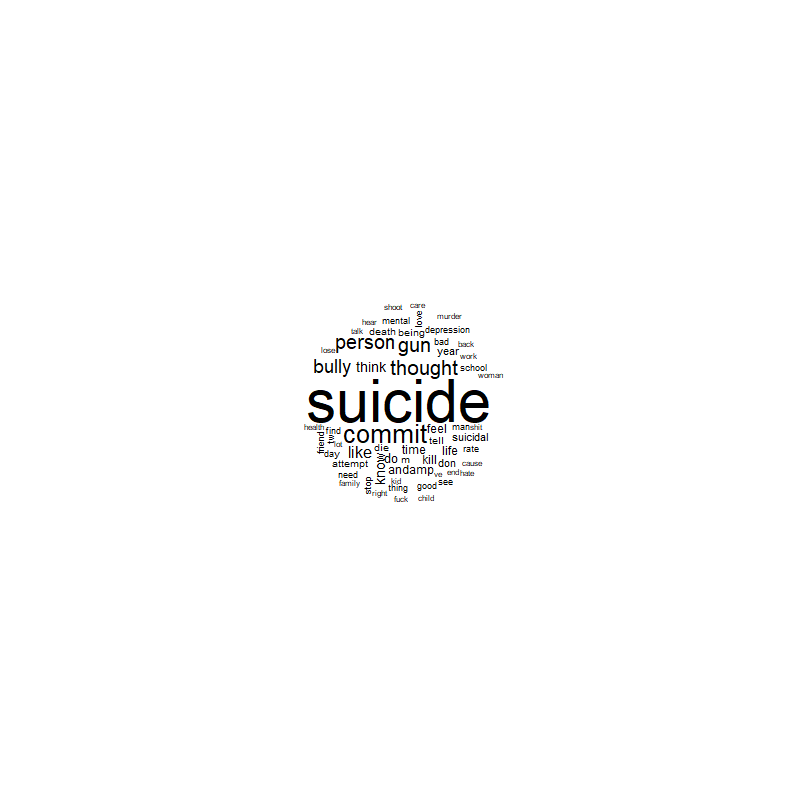


Figure 4 : Nuage de mots formes actives - Twitter - TS

Figure 5 : Nuage de mots formes actives - Twitter – non TS

L’analyse des similitudes des publications de Twitter liées au suicide montre une famille de mots principaux 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.

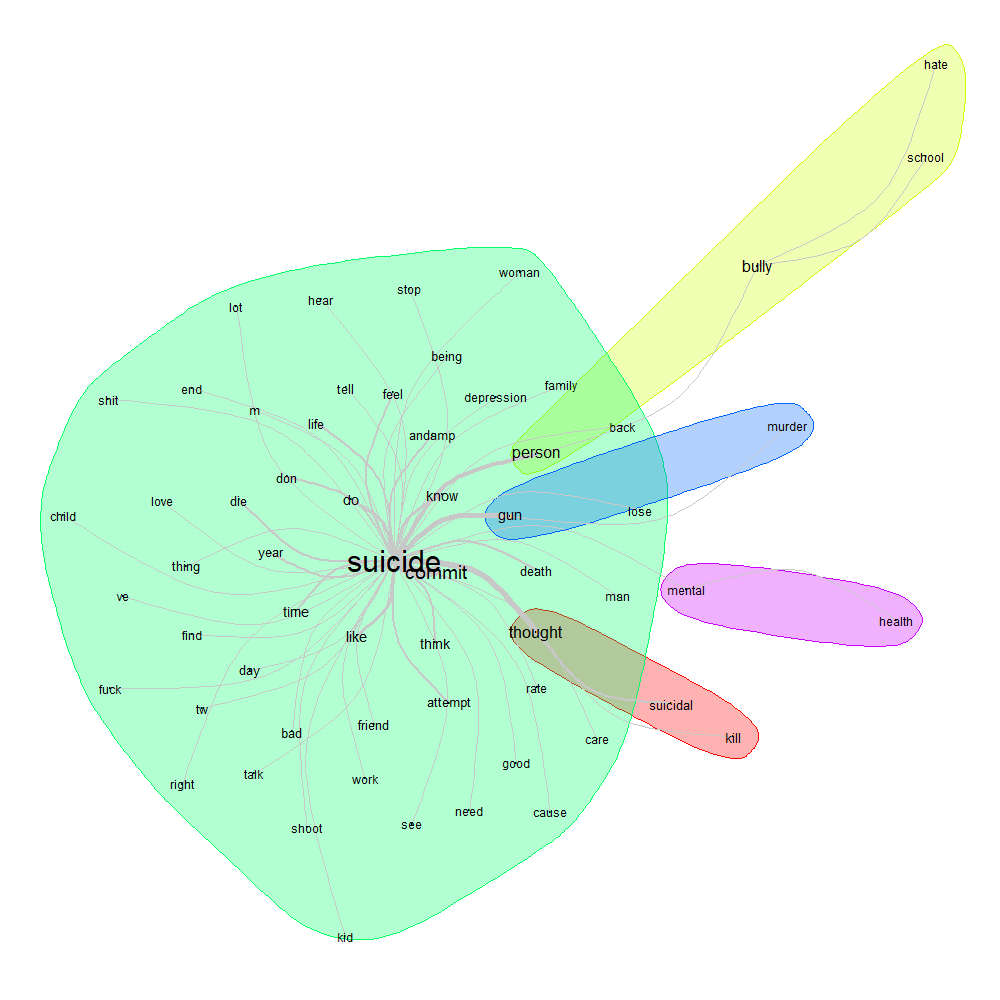


Figure 6 : Analyse des similitudes - Twitter - TS

Dans les données provenant de Reddit, on remarque une similitude dans les vocabulaires des deux catégories : les messages sont centrés sur les ressentis de chacun et sur leurs émotions.

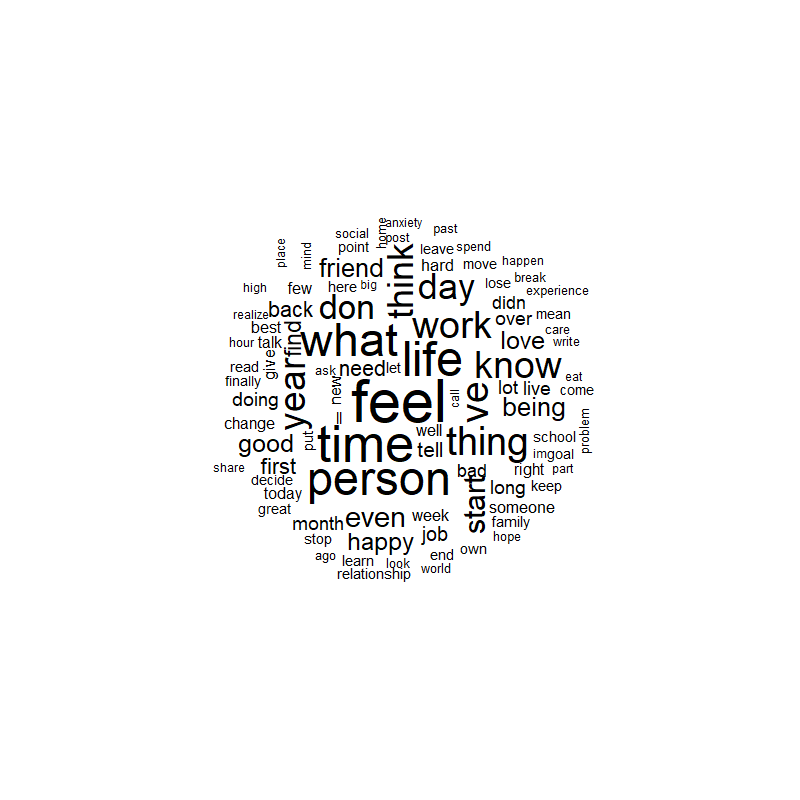


Figure 7 : Nuage de mots formes actives - Reddit - TS

Figure 8 : Nuage de mot formes actives - Reddit – non TS

D’ailleurs, si on se concentre sur les verbes les plus utilisés dans les deux catégories de Reddit on trouve feel (ressentir) et pratiquement les mêmes verbes.

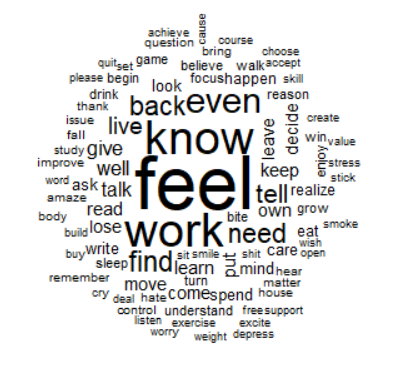
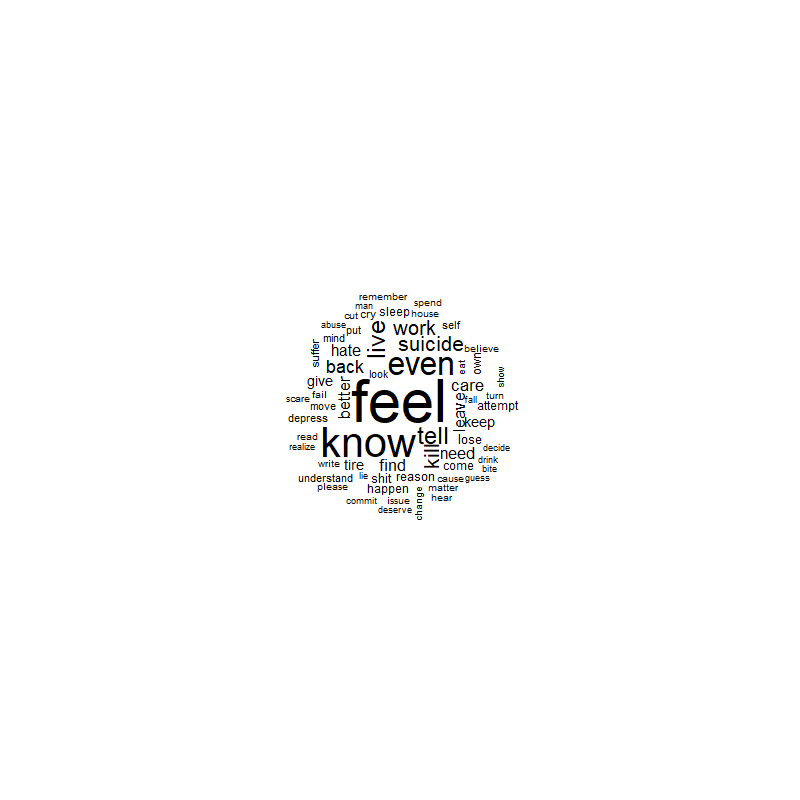


Figure 9 : Nuage de mots verbes - Reddit – TS

Figure 10 : 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 TS de Reddit.

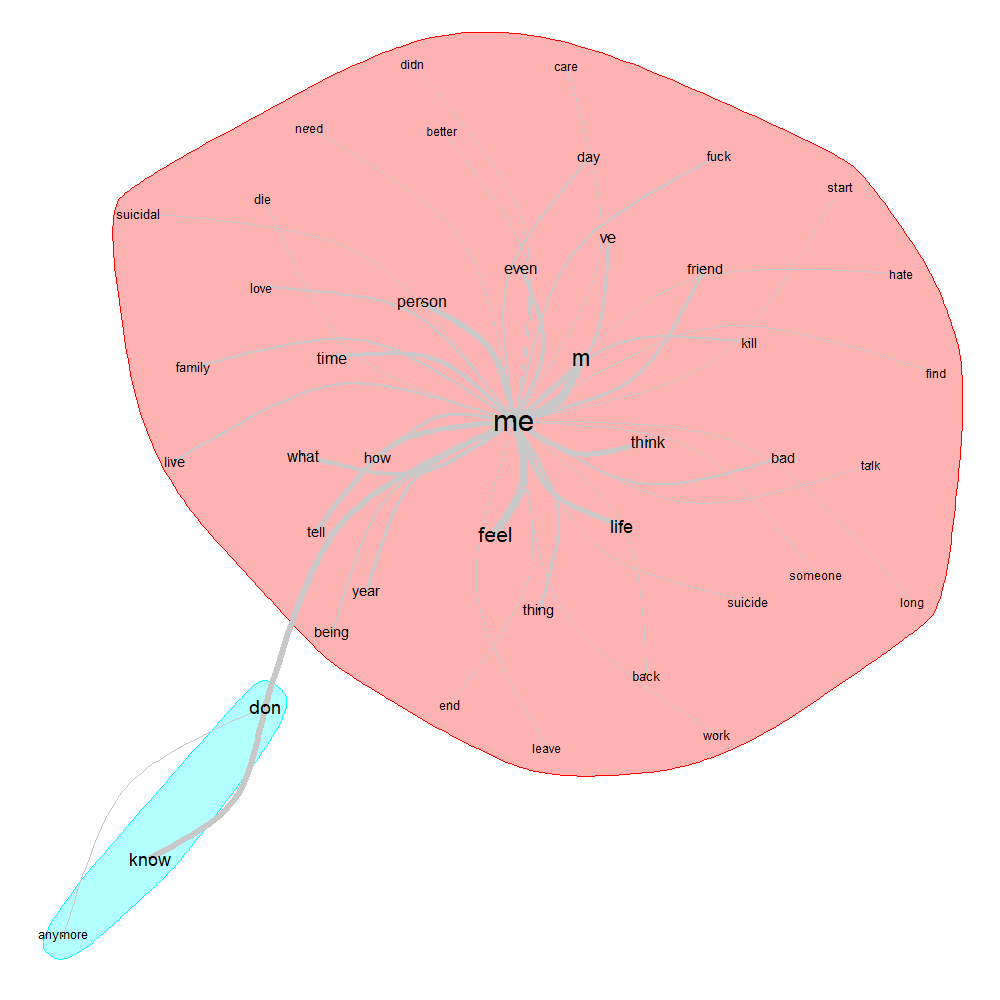


Figure 11 : Analyse des similitudes - Reddit - TS

En conclusion, nous utilisons trois jeux de textes assez différents, plutôt formel et axé sur le parcours de vie dans le cadre de Wikipédia, des messages courts et personnels pour Twitter et des messages assez longs et centrés sur les ressentis de ses auteurs pour Reddit. Cette diversité nous permet de faire des analyses sur des textes très différents et nous permet de faire des analyses plus complètes.