Compte rendu de réunion

Le 12/10/2021 à 12h-13h à l'université de Paul Valéry

Objet : Première exploration des données

Ordre du jour :

- Avoir plus d'informations sur les données
- Présenter notre première exploration des données
- Obtenir des informations sur les attendus de ce projet

Légende du tableau ci-dessous : P = Présent, A = Absent, E = Excusé, NC = non concerné par la réunion

Liste des participants au projet			
Nom	**Email**	**Fonction**	**Présent**
Sandra BRINGAY	sandra.bringay@univ-montp3.fr	Encadrante pédagogique	P
Florian LOMBARDO	florian.lombardo@univ-montp3.fr	Autre encadrant pédagogique	A
Laura SENECAILLE	laura.senecaille@etu.univ-montp3.fr	Etudiante	P
Anamé ROUMY	aname.roumy@etu.univ-montp3.fr	Etudiante	P
Lisa BETEILLE	lisa.beteille@etu.univ-montp3.fr	Etudiante	P
Matéo CALSACY	mateo.calsacy@etu.univ-montp3.fr	Etudiant	P
Jean CHABANOL	jean.chabanol@etu.univ-montp3.fr	Etudiant	P
Célia TEYSSIER	celia.teyssier@etu.univ-montp3.fr	Etudiante	P

Sommaire

Précisions sur les données

Twitter :

Les données ne contiennent aucune information personnelle telle que le genre, l'âge ou le lieu d'où a été écrit le tweet. De plus, parmi tous les tweets, beaucoup de catégories ne correspondent pas à des messages liés à des TS. On trouve par exemple dans la catégorie « cut » beaucoup de messages liés à l'action de se couper les cheveux. De plus, nous avons besoin de données non liées aux TS.

⇨ Pour les données personnelles, il n'est pas possible de les obtenir. Pour ce qui est des tweets sans rapport avec le suicide : ne pas conserver dans le jeu de données les catégories posant problèmes.

Wikipédia :

L'algorithme ne fonctionne plus, à nous de le reprendre pour le faire fonctionner. De plus, il nous faut trouver des catégories qui n'ont rien à voir avec le suicide pour la prédiction de TS.

Reddit :

Vous ne nous avez pas encore fourni les données liées à la catégorie *survivor*. Nous devons vous envoyer des noms de catégories sans lien avec le suicide pour que vous puissiez extraire les messages et nous les fournir.

Présentation de ce qu'on a fait

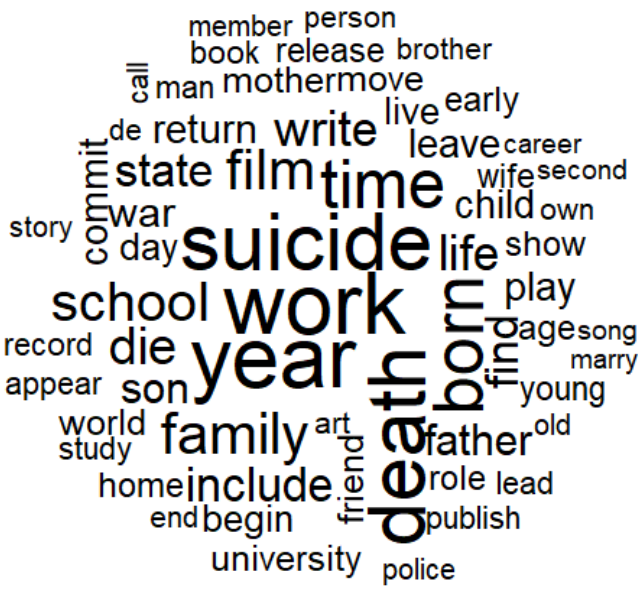
Comme convenu pendant la réunion précédente, nous avons commencé par travailler sur les données wikipédia. Nous avons importé les biographies de la catégorie « Death_by_suicide ».

Au total, après suppression des doublons et nettoyage des données à l'aide R et d'Iramuteq nous avons obtenu les informations suivantes :

- Nombre de biographies : 869
- Nombre de segments de texte, c'est-à-dire de suites de mots terminées par un délimiteur (ex : « . », « , », « ? », « ! », …) : 6 260
- Nombre total de mots (occurrences) : 218 426
- Nombre de mots différents (formes) : 23 785
- Nombre de mots qui n'apparaissent qu'une seule fois (hapax) : 12 411, c'est 5% des occurrences et 52% des formes.

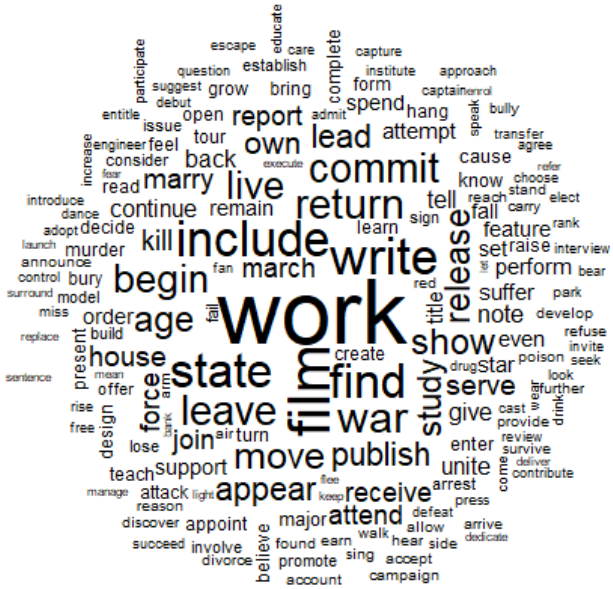
Le nombre d'hapax est assez élevé. Cela s'explique en partie par le fait que les biographies sont composées de beaucoup de noms propres comme des noms de personnes ou de lieux. Néanmoins, retravailler le nettoyage des textes pourrait aussi permettre de réduire ce nombre.

Le nuage de mot ci-dessous montre les formes les plus utilisées, toutes formes actives comprises (expression formes actives est à mettre en opposition avec l'expression formes supplémentaires, qui sont par exemple les pronoms, les conjonctions, etc.). Sans surprise, suicide fait partie des formes les plus utilisées, ainsi que year, born,… On comprend clairement que les biographies de Wikipédia sont assez standardisées et racontent la vie d'une personne du début à la fin de sa vie.

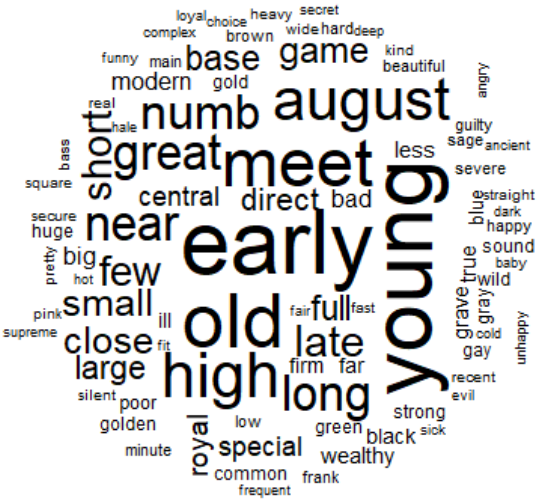


Le nuage de mot suivant montre uniquement les verbes les plus utilisés. On trouve notamment les formes work, school, write et film qui montrent une certaine focalisation sur le métier des individus et plus généralement sur leur parcours de vie. On peut d'ores et déjà remarquer que les biographies extraites de Wikipédia semblent appartenir à des profils plutôt artistiques.

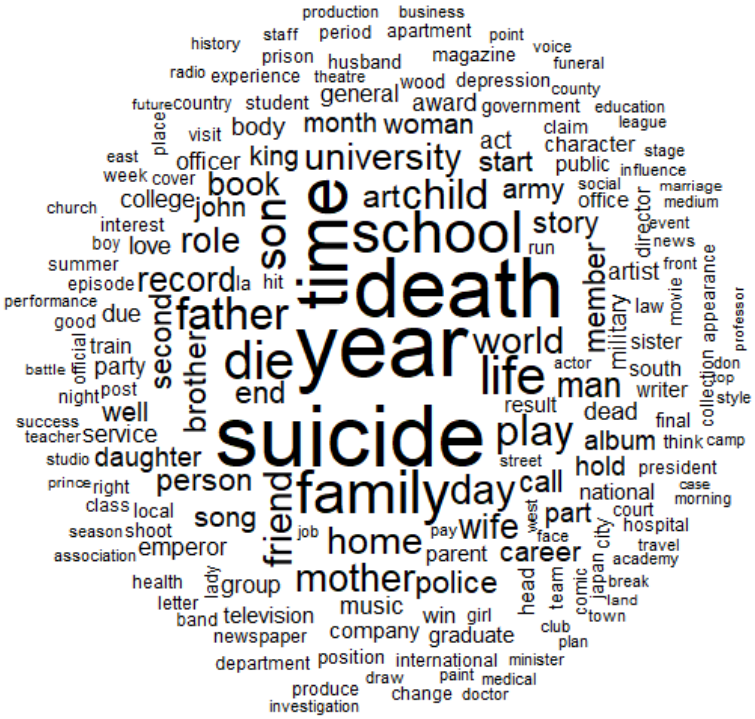
Néanmoins, les forme war, commit et attack semblent être beaucoup plus liées à l'action de se suicider.



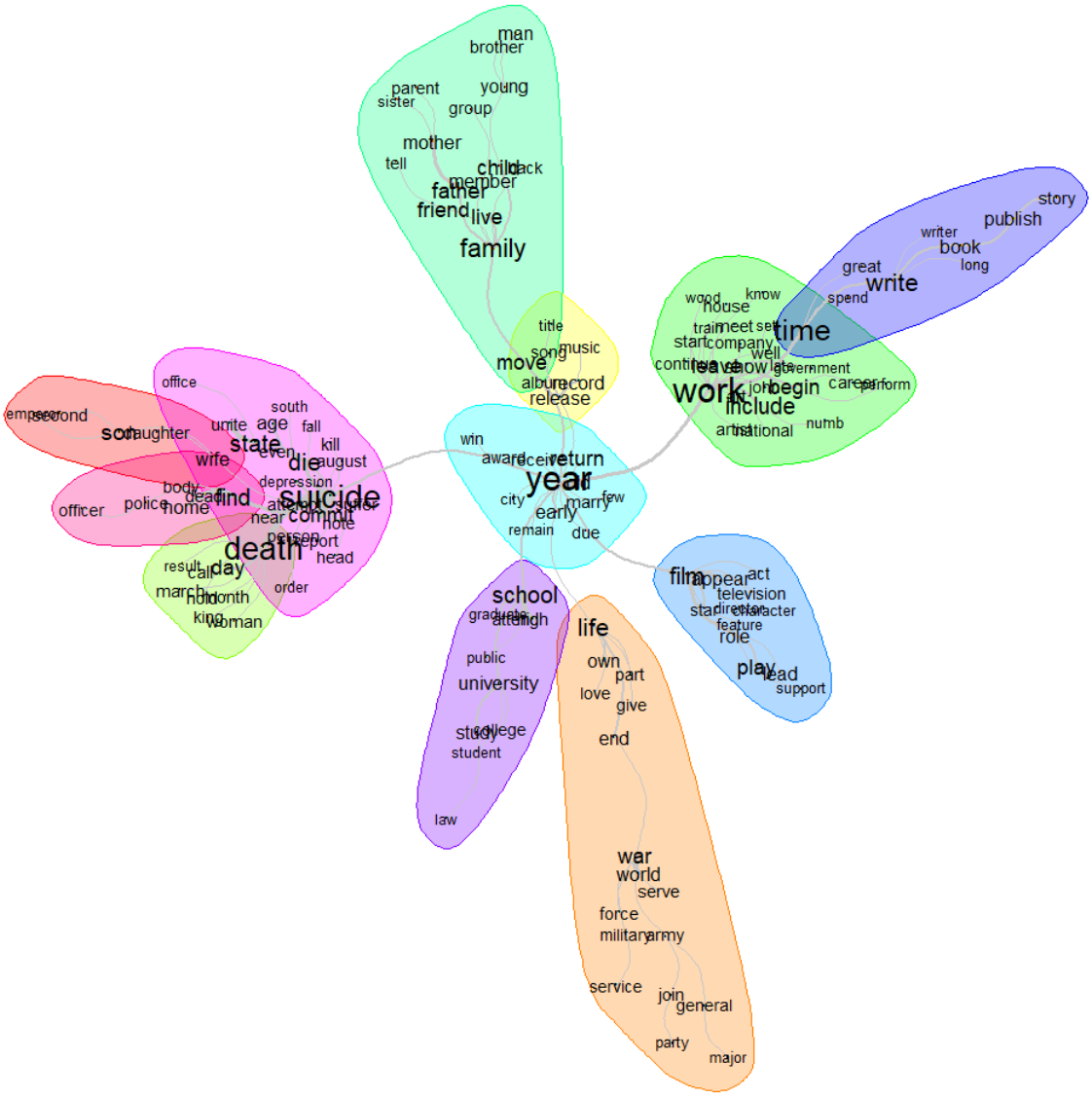
Le nuage de mot suivant résume les adjectifs les plus utilisés. Encore une fois, les formes utilisées semblent indiquer que les biographies sont écrites de façon plutôt standardisées. On peut tout de même remarquer que l'adjectif bad est plus utilisé que happy.



Enfin, le dernier nuage de mot est centré sur les noms les plus utilisés. La forme year fait partie des plus utilisées, ce n'est pas très surprenant puisque les biographie comportent souvent des phrases telles que « Né à… », « … à l'âge de … », etc. Parmi les noms, les notions de famille, d'école et de carrière sont très présentes. On peut supposer qu'elles ne sont pas liées à la notion de suicide contrairement aux formes telles que war, police, depression, prison, invastigation, etc.



Le prochain graphique montre une analyse des similitudes. On voit à nouveau clairement la notion de famille se dessiner La notion d'années qui est centrale. Cela n'a tien d'étonnant puisque l'âge de chaque personne est précisée plusieurs fois. Idem pour le mari ou la femme de chaque personne décrite. La notion de famille est proche de celle de la musique. La notion de travailler est proche avec celle d'écrire des livres. On trouve aussi une notion de films et tout ce qui y est lié. La notion de guerre est associée à l'amour, au début de la vie et à son terme. On trouve aussi une notion d'écoles/d'universités. Enfin, la notion de suicide est entourée par plusieurs catégorie, vraisemblablement ceux qui trouvent les victimes, donc la famille (fille fils femme) ou la police. La dernière catégorie proche de celle de suicide et laisse supposer que ce sont les conséquences du suicide : result, call, march.



Soutenance du TER

- Faire un plan détaillé pour savoir ce qu'on a à faire avant la présentation.
- Construire un jeu de données équilibré pour les 3 types de données.
- Faire une exploration des données. (Celles qui ont été établies pour wikipédia à appliquer sur les autres).
- Faire des modèles simples de classification, exemple : Random Forest ou SVM.
- Faire attention lorsqu'on cite des articles et utiliser Zotero pour générer des bibtex qu'on pourra mettre dans l'overleaf.

Résumé du travail fait pendant les années précédentes

L'année dernière :

L'objectif était de connaître l'humeur des gens avant et après la pandémie. Cela s'est avéré compliqué puisque les conditions d'utilisations pour extraire les tweets ont évoluées en cours d'année et ont limité la période d'extraction à une semaine.

Il y a deux ans :

Le groupe de TER cherchait s'il était possible de construire une trajectoire dans les biographies Wikipédia pour retrouver des évènements similaires chez les individus qui se sont suicidés.

Pour la fin du semestre

Les objectifs pour la fin du semestre sont :

- Avoir tous les jeux de données (Wikipédia, Reddit et Twitter) en rapport et non en rapport avec le suicide.
- Explorer les jeux de données avec Iramuteq et faire des comparaisons entre les jeux de données et les catégories suicide vs les autres.
- Faire des modèles de prédiction de « base ».
- Faire un état de l'art de 2 à 3 pages pour raconter ce que les chercheurs ont faits avant nous. Cette partie comprendra une partie plutôt sociologique et une autre sur les techniques de prédiction.

Actions à entreprendre

- Fournir les catégories à importer pour Reddit et les mots clés pour extraire les tweets sans rapport avec le suicide.
- Bien faire les explorations de données pour les catégories en lien avec le suicide et celles qui ne le sont pas.
- Améliorer le nettoyage des jeux de données.