Compte rendu 1 : Question(s) de recherche :

-- Questions posées aux commanditaires :

- Qui est-il/elle? où travaille t-il? Quel est son rôle dans sa «compagnie»?
  - Sandra Bringay est la directrice de l'UFR 6 MIAP à l'université Paul Valéry Montpellier III. Elle enseigne également des modules en Licence et Master MIASHS. En parallèle, elle effectue des recherches aux LIRMM.
    - sandra.bringay@univ-montp3.fr
  - Florian Lombardo, en thèse de sociologie et ~~temporairement~~ enseignant.
    - florian.lombardo@univ-mont p3.fr
- Pourquoi le problème est-il important pour il/elle, pour son entité, pour la société?
  - Ce problème est important car cela permettra de déceler les comportements suicidaires à travers les ~~posts~~ sur les réseaux sociaux. Énormément de personnes font des tentatives de suicides et quasiment aucun suivi n'existent pour les contrer.
  - Malheureusement, de nos jours il n'y a pas assez de médiatisation sur le suicide ou de démarche publique. De plus, rien ne nous permet de prévoir à l'avance si quelqu'un va se suicider. L'intérêt de ce sujet est donc de pouvoir identifier, par une analyse textuelle, les personnes étant les plus à risque de faire une tentative de suicide pour ainsi sensibiliser, concrétement et rapidemment.
  - Autre formulation : Chercher à comprendre les comportements suicidaires, et s'il y a un moyen de les prédire.
- Qu'est-ce qui vous est demandé exactement de résoudre?
  - Mission : Elaboration d'un algorithme de prédiction des risques suicidaires à partir de corpus de textes
- En quoi ce sujet est-il intéressant ?
  - Le suicide représente la cause de mortalité la plus importante chez les jeunes.
    - Pour les 16-25 ans : deuxième cause de mortalité après accident de la route
    - Pour les 25-39 ans : première cause de mortalité
  - Le suicide est lié a de nombreux facteurs sociaux (exclusion, repli sur soi, problème professionnel ou personnels, dépression…). Donc il serait intéressant de comprendre les états des personnes qui feraient des TS.
- Pourquoi pensez-vous que l'on vous demande de résoudre ce problème en particulier? Seriez-vous intéressé à résoudre cela plutôt?
  - Ce problème traite un sujet très important, il est important de chercher à le résoudre pour "essayer de réduire" les risques associés et de les comprendre.
- Quel est le contexte actuel dans lequel on vous demande de résoudre ce problème? Pourquoi est-ce important de résoudre ce problème maintenant?
  - Le suicide est la 2ème cause de mortalité chez les jeunes.
  - **Contexte :** Une personne met fin à ses jours toutes les 40 secondes dans le monde. Le suicide est la 2e cause de mortalité des 15-29 ans en France. Parfois un simple contact humain suffit pour l'éviter. Le défi réside dans la

prédiction des comportements suicidaires. Elle peut se faire par d'une part de l'analyse des contenus textuels publiés sur les réseaux sociaux.

- La finalité ?
  - Prévention suicidaire. Création d'un algorithme qui détecte un tweet à risque et envoie une notification → permet d'avoir un suivi médical, avec un but d'une mise en contact avec des groupes de survivants suicidaires pour apporter un soutien émotionnel.
- Ce qui a déjà été fait :
  - **Conclusion du projet l'année dernière:** Algo qui arrivait à prédire si un corpus de texte faisait référence au suicides au travers des émotions des personnes (uniquement sur Reddit).
    → Algo des M2: variables pour représenter les émotions dans les textes pas très efficaces.

-- Reformuler le problème de façon non technique et valider cette formulation auprès du commanditaire:

- Traduire leur requête ambiguë en un problème concret et bien défini.
  - **Notre projet pour cette année sera donc de mettre en place** un algorithme qui identifie les textes présentant un risque de tentative suicidaire à partir de toutes les données dont on dispose (données tirées de Reddit, Wikipédia et Twitter). Cet algorithme devra prendre en compte toutes ces données, mais il aura une performance différente en fonction de la provenance de celles-ci. Cet algorithme devra, selon la source et le format des données, s'adapter pour être le plus efficace possible.

    Notre travail devra également s'appuyer sur les travaux des années précédentes. Dans un premier temps, les corpus de texte analysés feront référence uniquement au suicide et non au mal être d'une personne (texte traitant des idées noires par exemple). Mais il serait intéressant pour la finalité de pouvoir analyser également ces mal êtres.

- Pouvez-vous expliquer simplement le problème à quelqu'un d'autre?
  - Prédire les comportements suicidaires au travers d'algorithmes de prédilection.
  - Le résultat de notre projet pourra être présenté sous la forme d'un site internet pour avoir une présentation plus soignée et plus claire.
- Formuler des questions qui définissent le «business problem» et qui peuvent être attaquées par une technique de data science (prédiction, classification, clustering, recommandation, modélisation statistique, etc.)
  - Utilisation de classification supervisée et non supervisée.

Compte rendu 2 : Planification et gestion de projet :

- **Team Lead** : Anamé Roumy.

- **Diagramme de GANTT** : Pour l'élaboration du diagramme de GANTT nous avons utilisé Excel. Nous avons imaginé un agenda possible tout au long du semestre en fonction de la durée et de la difficulté des étapes. Nous avons essayé de répartir les tâches sur les semaines de cours en prévoyant de les poursuivre en semaine d'alternance si nécessaire.
  Chaque couleur regroupe les étapes dépendantes les unes des autres. Par exemple, le nettoyage des données se fera une fois l'étape "Parcourir les données et visualisations simple de données" réalisée.

Semaines en entreprises														
Septembre			Octobre				Novembre				Décembre			
13/09 - 19/09	20/09 - 26/09	27/09 - 03/10	04/10 - 10/10	11/10 - 17/10	18/10 - 24/10	25/10 - 31/10	01/11 - 07/11	08/11 - 14/11	15/11 - 21/11	22/11 - 28/11	29/11 - 05/12	06/12 - 12/12	13/12 - 19/12
Semaine 2	Semaine 3	Semaine 4	Semaine 1	Semaine 2	Semaine 3	Semaine 4	Semaine 1	Semaine 2	Semaine 3	Semaine 4	Semaine 1	Semaine 2	Semaine 3
Création et connexion GitHub Mise en place des outils de colab (Trello, Discord, Watsapp)													
	Bibliographie Résumé Parcourir les données et Visualisations simples de données			Bibliographie Résumé Nettoyage des données				Conception et test des différents modèles de prédiction Site web simple (avec présentation projet + graphique) PPT + RENDUS				Préparation Oral	Anglais

- **Outils de gestion de projet** : Nous avons choisi d'utiliser Trello pour la gestion de notre projet car certains d'entre nous avons eu l'habitude de travailler sur cet outil. De plus, il est très facile d'utilisation, accessible facilement et intuitif. Ajouté à celà, nous allons exploiter les fonctionnalités de Git pour le partage de documents et de codes. Enfin, pour le travail collaboratif nous nous appuierons sur les outils Discord et Whatsapp Notre stratégie est de faire participer chaque membre du groupe aux étapes.

- **Stratégie définie** : Pour la bibliographie nous lirons chacun un texte afin de pouvoir acquérir quelques connaissances sur le sujet que nous mettrons en commun à la fin de notre lecture. En ce qui concerne les autres tâches, pour une efficacité plus poussée, nous désignerons les "sous-tâches" à chacun des membres en fonction de leurs compétences et de leurs appétences. Nous travaillerons par 1, 2 ou 3 selon la difficulté de la sous-tâche.

Articles	Membres du groupe
Detection of Suicide Ideation in Social Media Forums Using Deep Learning -- *Michael Mesfin Tadesse , Hongfei Lin, Bo Xu, Liang Yang*	Célia TEYSSIER & Matéo CALSACY
The integrated motivational – volitional model of suicidal behaviour -- *Rory C. O'Connor, Olivia J. Kirtley*	Jean CHABANOL & Anamé ROUMY

Unveiling Online Suicide Behavior: What Can We Learn About Mental Health from Suicide Survivors of Reddit? -- *Ashwin Karthik Ambalavan, Bilel Moulahi, Jérome Azé, Sandra Bringay*	Laura SENECAILLE
The Interpersonal Theory of Suicide -- *Van Orden et al.*	Lisa BETEILLE

+ Autres articles cf TER 2020 qui traite de **TWITTER**

Compte rendu 5 : Outils et programmes à utiliser :

- Outils permettant d'optimiser l'efficacité du travail et qui ont également permis d'accroître notre productivité :
    - **Google Drive** : le Google Drive nous a permis de nous organiser pour les rédactions de comptes rendus, les résumés d'articles, les agenda minutes etc… Nous pouvions suivre en temps réel ce que chacun écrivait pour donner un avis ou modifier des choses directement sur la même page, celà nous permettait d'optimiser notre travail d'être plus efficace.
    - **Github** : le Git nous a permis de nous partager les données mise à notre disposition (tweets, reddit, wikipédia, articles …) ainsi que les codes de classification plus rapidement, et ainsi pouvoir avoir tous les mêmes données pour travailler ou pour voir ce que les autres ont ajouté. Nous avons également utilisé Github pour créer une "to do list" avec les éléments à faire ("to do"), en cours ("in progress") et terminés ("done"), cela nous permettait de rester organiser, d'avoir une vue d'ensemble de tout ce que nous devions faire et d'attribuer des tâches à chaque membre pour être plus efficace.
    - **Discord et Whatsapp** : nous avons pu optimiser notre travail grâce à ces outils de communication, car cela nous permettait de rester en contact pendant nos périodes d'alternance ou le weekend pour rester productifs. Discord nous permettait d'entretenir des réunions régulièrement en appels vocaux et Whatsapp pour des messages réguliers plus instantanés. Discord nous a aussi permis d'entrer en contact avec Mathéo Daly et Florian Lombardo plus rapidement lorsque nous avions des soucis dans notre projet.

Tous ces outils étaient compatibles avec entre les différents OS, sauf pour Github où nous avons eu des soucis avec les ordinateurs Macbook qui créaient des dossiers DS_Store dans les fichiers, nous obtenions donc des erreurs au moment de push et de pull. Nous avons pu néanmoins régler ce souci en utilisant un gitignore.

- Outils collaboratifs choisis pour ce projet :
    - **Discord** pour la communication entre les membres du projet en appels vocaux avec des partages d'écran.
    - **Google Drive** pour le partage de documents ainsi que pour faire des rédactions collaboratives.
    - **GitHub** pour le partage de répertoires (codes, articles, données…) ainsi que pour la "to do list".
    - **Whatsapp** pour une communication plus rapide que Discord.
    - **Overleaf** pour la rédaction du rapport final en LaTeX.
    - **Google Colab** pour le partage des codes de classification entre nous, pour voir ce que nous obtenions en faisant marcher ces codes avant de les push sur Github.

- Langage de programmation :
  Utilisation du **R** pour le nettoyage utile à l'exploration des données. Nous avons privilégié R pour le nettoyage car Célia avait plus l'habitude de travailler avec cet outil, nous allons donc être plus efficaces dessus. De plus, il possède des packages très utiles pour gérer rapidement des gros volumes de données.

Iramuteq pour réaliser l'exploration des données. Iramuteq est un logiciel basé sur R créé par un groupe de chercheurs et il est très utile pour visualiser et analyser des textes, c'est pour cela que nous avons choisi de travailler dessus.
Python pour la faire la classification, nous avons pu créer des modèles de prédiction (SVM, Random forest et Kmeans). Nous avons choisi Python pour réaliser la classification car c'était ce qui était le plus adapté et préconisé par nos commanditaires.

Reproductible veut dire ici peut être reproduit, comparé ?

- Approche reproductible ?
  Notre approche peut être reproductible car nous avons su bien nous répartir le travail, utiliser des outils qui nous ont permis d'avoir tous accès aux mêmes informations rapidement et de garder une communication régulière même pendant nos périodes d'alternance. Notre approche nous a donc permis d'être efficace et de garder une bonne entente générale dans notre groupe, ce qui est et sera important pour mener à bien ce projet.

Hors sujet

1.

- Approche permettant de garder une trace des modifications ainsi que de stocker et rendre plus accessible les données :
  L'utilisation de **GitHub** nous permet de garder une trace des modifications et des échanges de répertoires car nous avons tous le même code à partir du moment où nous le récupérons sur le Git par conséquent rien ne peut être perdu.
  L'utilisation de **Google Drive** nous a également permis de garder une trace de nos modifications car nous avions accès en temps réel aux modifications des fichiers établies par tous les membres du groupe. De plus, ces modifications sont sauvegardées instantanément, donc nous ne perdons rien. De même pour l'utilisation de **Google Colab**, qui nous a permis de garder toutes les traces de modifications de codes que nous avions réalisées.