Compte rendu de réunion

Le 08/12/2021 à 15h30-16h en visio sur Discord

Objet : Première exploration des données

Ordre du jour :

- Présenter des classifications sur les données Reddit
- Avoir Récupéré toutes les données manquantes pour Wikipédia et Twitter
- Avoir finis de rédiger des comptes rendus

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	P
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

Avancées

- Nous avons pu régler le problème de l'algorithme de Wikipédia et récupérer le document csv contenant les données des suicides de 1920.
- Plusieurs éléments de classifications ont été réalisé sur les données Reddit.
- Avancement de l'état de l'art par Matéo et Jean
- Nous avons reçu les données « happy » de twitter que Sandra nous a envoyé.

Difficultés rencontrées

- Nous avions rencontré encore des problèmes avec l'algorithme de Wikipédia qui ne marchait pas sur nos ordinateurs, mais nous avons réussi à le faire marcher.
- Célia ayant eu le covid, cela nous a mis en retard pour travailler sur les données et nous avons du repousser la classification sur les données twitter.
- Nous avons rencontré des problèmes sur la classification.

Plan d'organisation

- Pour plus d'efficacité, nous nous sommes réparti le travail pour la semaine à venir:
  - Matéo et Jean doivent finir l'état de l'art et le compte rendu étape 4.
  - Célia devra faire de la visualisation.
  - Laura et Anamé devront continuer et finir la classification.
  - Lisa s'occupera de la rédaction du compte rendu 5, de la gestion de projet et de la correction des étapes (voir les mails de correction de comptes rendus envoyés par Sandra).

Actions à entreprendre ou éléments de corrections à apporter

- Pour la classification, il faut qu'on fasse attention à ce qu'il n'y ait pas de titres dans les mots qui se répètent. Nous utilisons seulement 100 données pour Ask Reddit, ce n'est pas assez, il faut changer de catégorie car des classes trop déséquilibrées ne font pas de bons modèles. Il faut aussi que nous fassions attention à ce qu'on affiche bien le nombre d'occurence et pas le tfidf. La matrice de confusion donne l'impression que l'on a fait de la prédiction sur des données différentes (1000 données).

  - Dans la partie du notebook avec les csv, penser à supprimer en amont des mots qui peuvent fausser nos résultats (comme « don't »)
  - Pour la partie sur le SVM, il faut réussir à lancer un autre classifier car il peut exister plusieurs algorithme de classification suivant les données, ce n'est pas le même qui va marcher le mieux en fonction de ces données.

- Le tableau que l'on doit obtenir our chaque type de données :

	Rappel	Précision	1 F
Modèle 1			
Modèle 2			
…			

- Envoyer en urgence les étapes à Sandra pour qu'elle nous annote des éléments de correction qu'on devra modifier avant de les renvoyer pour la notation.
- Mercredi : répétition finale avec Sandra
  - Finir de rédiger les documents
  - Faire ce qu'on a fait avec la méthode supervisée sur le forum Reddit mais sur les données twitter et inversement.
  - Rajouter un ou deux classifier