

Compte rendu 5 : Outils et programmes à utiliser

Introduction

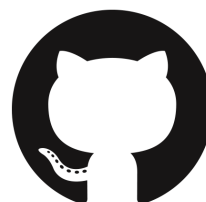
Pour la réalisation de ce projet il a fallu choisir de bon outils et programmes à utiliser pour ne pas nous ralentir et pour nous faciliter les tâches.

Outils et programmes à utiliser

Pour ce projet, nous avons réfléchi à la mise en place d'outils collaboratifs pour nous permettre d'optimiser l'efficacité de notre travail et ainsi d'accroître notre productivité. Pour l'organisation des rédactions de comptes rendus, de résumés d'articles et des agendas minutes nous avons utilisé Google Drive. Cet outil nous a permis d'optimiser notre travail et d'être plus efficace en suivant en temps réel ce que chacun écrivait pour donner un avis ou modifier des choses directement sur la même page. De même pour le Google Colab, qui a permis de partager nos codes de classification entre nous après les avoir terminés sur Jupyter Notebook. Cela nous permet de voir ce que nous obtenons en faisant fonctionner ces codes avant de les push sur Github. De plus, nous avons relié directement les données du Github au Google Colab pour nous faciliter le travail.

Le Github nous a servi à partager les données mises à notre disposition (tweets, reddit, wikipédia, articles ...) ainsi que les codes de classification plus rapidement, et par conséquent pouvoir avoir tous les mêmes données pour travailler ou pour voir ce que les autres ont ajouté. Nous avons aussi utilisé Github pour créer une "to do list" avec les éléments à faire ("to do"), en cours ("in progress") et terminés ("done"). Cela nous permet de rester organisé, d'avoir une vue d'ensemble de tout ce que nous devons faire et d'attribuer des tâches à chaque membre pour être plus efficace.

Notre travail a notamment pu être optimisé grâce à Discord et Whatsapp. Ces outils nous ont permis de rester en contact pendant nos périodes d'alternance ou les weekends pour rester productifs. Nous avons utilisé Discord pour entretenir des réunions régulièrement en appels vocaux en utilisant des partages d'écran, et Whatsapp pour des messages réguliers plus instantanés. Discord nous a également servi pour entrer en contact avec Mathéo Daly et Florian Lombardi plus rapidement lorsque nous avons des soucis dans notre projet (notamment notre problème d'algorithme Wikipédia).



Nous avons utilisé plusieurs langages de programmation pour mener à bien notre projet. Il a été important d'utiliser RStudio pour le nettoyage utile à l'exploration des données. Nous avons privilégié RStudio 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 nous a servi pour réaliser l'exploration des données. Ce logiciel basé sur RStudio a été créé par un groupe de chercheurs. Il est très utile pour visualiser et analyser des textes, c'est pour cela que nous avons choisi de travailler dessus. Pour finir, Python nous a été très utile pour faire la classification. Nous avons pu créer des modèles de prédictions comme des 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.

Une approche reproductible signifie une approche qui peut être reproduite ou copiée. L'approche que l'on a utilisé pour notre projet peut être qualifiée de reproductible car nous utilisons des notebooks pour écrire nos codes, ce qui permet de changer le nom des données à utiliser pour pouvoir faire marcher tout le code. Par exemple, le code pour le Kmean sur Reddit peut juste être remplacé par les données de Wikipédia pour que ça marche sans rien changer au code.

Il était important d'adopter une approche permettant de garder une trace de toutes les modifications apportées ainsi que de stocker et rendre plus accessible les données. Pour cela, en utilisant le GitHub, cela 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. Grâce au Google Drive, cela nous a également permis de garder une trace de toutes nos modifications car nous avons accès en temps réel aux changements des fichiers établies par tous les membres du groupe. De plus, ces changements sont sauvegardés instantanément, donc nous ne perdons rien non plus. De même pour l'utilisation de Google Colab qui nous a permis de garder toutes les traces de modifications de codes que nous avons réalisées.

Difficultés rencontrées avec ces outils

Tous ces outils étaient compatibles 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 obtenons donc des erreurs au moment du push et du pull. Nous avons pu néanmoins régler ce souci en utilisant un gitignore.

Conclusion

Ces outils et programmes nous ont permis d'optimiser notre travail et de nous rendre plus efficaces et plus productifs. Les connaissances de certains membres de l'équipe sur certains programmes nous ont également été bénéfiques pour notre productivité et notre organisation.

Bibliographie

<https://www.cnrtl.fr/definition/reproductible>