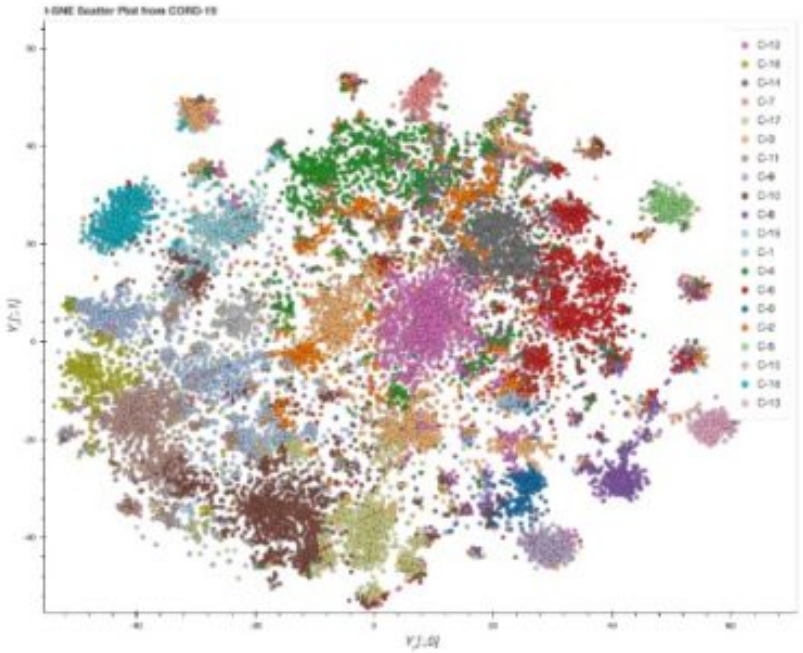
Retour sur le challenge CORD-19 et la plateforme Kaggle

Nous vous en parlions dans la précédente newsletter, le challenge CORD-19 lancé par la Maison Blanche et son Office of Science and Technology Policy (OSTP) avait pour but l'exploitation d'un grand volume de données publiées issues de la recherche afin d'accélérer encore la recherche contre le virus. [1, 2]

Après un peu plus d'un mois de lancement, plus de 1200 contributions ont été faites à ce défi lancé permettant d'analyser le corpus de recherche et faciliter la réutilisation de ces informations. Parmi les soumissions, certaines sont de simples fonctions permettant de facilement extraire les informations de base (auteurs et leurs affiliations universitaires, titre de l'article, abstract, corps de l'article etc.) alors que d'autres sont des solutions plus complètes faisant en particulier appel à des algorithmes d'apprentissage non-supervisé (notamment méthodes de clustering), des techniques de traitement automatisé du langage naturel (dites Natural Language Processing ou NLP) ou dans certains cas des solutions qui vont jusqu'à proposer une interface graphique.

COVID-19 Literature Clustering

A ce jour, la contribution la plus populaire a été produite par une équipe du **Malware Research Group** de l'Université de Maryland à Baltimore County. L'algorithme proposé permet de grouper et visualiser les articles scientifiques pertinents en fonction de mots-clés fournis par l'utilisateur (voir ci-dessous).



Regroupement d'articles scientifiques sur le coronavirus par mots-clés – illustration tirée de
Kaggle. Chaque point sur le graphique représente un article de recherche. En cliquant sur un
point, on obtient les informations (titre, auteurs etc.) et un lien vers l'article. A chaque groupe
d'articles trouvé par l'algorithme, une couleur est associée.

L'interface complète est accessible au lien suivant [3].

Discovid.ai

Une autre solution, développée par un étudiant allemand au sein du groupe de recherche TECO à l'institut de technologie de Karlsruhe (KIT), a été transposée en un service web complet et mis à disposition sur un site dédié nommé Discovid.ai. Ce dernier permet de trouver à partir d'un article d'intérêt, les articles qui lui sont associés mais se présente plutôt sous la forme d'un moteur de recherche. La barre de recherche permet également d'entrer une question relative à la recherche contre le COVID-19 et être renvoyé vers les articles pertinents, avec un score de similarité associé correspondant. Diverses fonctions de filtre notamment en fonction de la date

de publication sont également disponibles [4, 5].

La contribution d'un lycéen indien

Une contribution a été réalisée par un jeune étudiant indien de 17 ans, Tarun Sriranga Paparaju [6, 7] et fait appel à une analyse par traitement automatique du langage naturel (NLP). Entre autre, la solution proposée permet de faire une étude des pays d'origine des auteurs, une « sentiment analysis » (détecte si un article utilise un vocabulaire plutôt positif, négatif ou neutre). Plus original encore, la proposition d'algorithme tente, à partir de différentes méthodes statistiques, d'identifier des pistes thérapeutiques sur la base de la littérature scientifique disponible. Si l'on se fie aux données présentées, le chemin emprunté permettrait de retomber sur l'hydroxychloroquine comme possible traitement à envisager mais également sur la doxorubicine (médicament anticancéreux qui lui ne semble pas être pertinent).

D'autres jeux de données partagés sur Kaggle

Outre le défi officiel lancé par la Maison Blanche, d'autres jeux de données complémentaires ont été rendus disponibles sur la plateforme Kaggle afin que tout data scientist puisse les manipuler et expérimenter leurs algorithmes.

Les données statistiques sur le nombre de cas confirmés par pays et le nombre de décès collectées par l'université Johns Hopkins ont notamment été rendues accessibles et donnent lieu à une compétition hebdomadaire pour affiner les algorithmes statistiques de prédiction qui permettraient d'aider les personnels soignants et décideurs politiques [8]. Un autre jeu de données a été publié par une équipe, **EpidemicForecasting.org**, de l'université d'Oxford au Royaume-Uni en collaboration avec l'Australian National University et l'université d'Harvard [9]. Dans ce jeu de données, on retrouve notamment une description textuelle des mesures d'endiguement de l'épidémie, les dates de début et de fin (lorsqu'elles sont

disponibles) ainsi que des mots-clés associés (tels que interdiction de voyage, avis de mise en isolement pour les personnes âgées etc.).

Une plateforme d'entraide et d'apprentissage

Si la pertinence des différentes contributions et de leur utilité avérée semble, à l'heure actuelle, relativement modeste, l'initiative a le mérite de permettre d'organiser la réflexion et de fédérer des communautés d'utilisateurs aux compétences et profils différents, réunies autour d'un problème commun. On notera l'exemple de **CoronaWhy** [9], initiative lancée sur la plateforme Kaggle, qui, face à l'exhaustivité du défi CORD-19 et au manque de coordination, a permis de structurer un groupe de plus de 900 personnes (experts techniques ou non techniques) à travers le monde pour avancer ensemble sur quatre des dix tâches initialement prévues.

Ces différents exemples témoignent de l'importance de l'entraide et il faut les envisager davantage d'un point de vue éducatif et pédagogique que d'un point de vue opérationnel.

Sources:

[1]
https://www.whitehouse.gov/briefings-statements/call-action-tech-community-new-machine-readable-covid-19-dataset/
[2] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks
[3]
https://maksimekin.github.io/COVID19-Literature-Clustering/plots/t-sne_covid-19_interactive.html
[4] https://www.kaggle.com/danielwolffram/topic-modeling-finding-related-articles
[5] https://discovid.ai/search
[6] https://srirangatarun.wordpress.com/home/
[7]

https://www.kaggle.com/tarunpaparaju/covid-19-dataset-gaining-actionable-insights
[8] https://www.kaggle.com/c/covid19-global-forecasting-week-4/data
[9] http://epidemicforecasting.org/containment
[10] https://www.coronawhy.org/