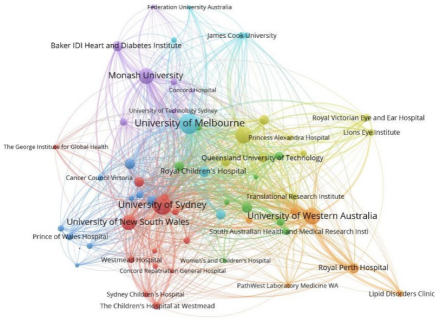
Construction et exploration d'un graphe de collaboration scientifique



Encadrant·e/Supervisor : Charlotte Laclau (S2A, IDS) et Mathide Perez (S2A, IDS) - prenom.nom@telecom-paris.fr
Nombre d'étudiant·es minimum dans chaque instance de ce projet 4
Nombre d'étudiant·es maximum dans chaque instance de ce projet: 5
Combien d'instances de ce projet proposez-vous ? 1
Tags : graphe, prédiction de lien, graphe de collaboration, science des données, Python

1 Contexte/Context

L'objectif de ce projet est de construire un graphe de collaboration scientifique entre des chercheurs et chercheuses en France. Dans ce graphe, la collaboration scientifique est établi si deux chercheurs ont co-écrit un article scientifique ensemble. Dans un deuxième temps, les étudiants feront une exploration de ce jeu de données (visualisation et statistiques élémentaires). Ce nouveau graphe pourra ensuite servir à la communauté scientifique comme nouveau jeu de données de référence pour l'étude de graphes temporels et la détection de biais dans les graphes temporels.

2 Attendus du projet/Expectations

Le point de départ pour les étudiants sera d'extraire du site HAL (https://hal.science/) les publications scientifiques de chercheurs et chercheuses en France et en extraire un graphe de collaboration. Pour ce faire, les étudiants pourront s'appuyer sur l'API du site HAL. Le périmètre géographique et temporel sera à définir avec les étudiants.

Le graphe ainsi construit devra stocker un certain nombre de méta-données tels que le domaine de recherche, les années de publications ou l'institution à laquelle le chercheur/la chercheuse appartient. Idéalement, les étudiants devront également proposer une approche pour extraire des informations supplémentaires à partir des prénom des chercheurs, telles que le genre.

Une fois ce graphe construit, il s'agit de proposer une première exploration (visualisation et statistique) à partir de librairies Python telles que scikit-nework ou networkx. Les étudiants devront donc effectuer une recherche bibliographique afin d'identifier les outils nécessaire (quelles statistiques, quelles méthodes de visualisation) et d'établir une procédure pour l'analyse du graphe construit. L'objectif de cette partie consiste essentiellement à comprendre et à caractériser la structure du graphe construit. Pour l'aspect temporel, on s'intéressera également à des outils permettant de caractériser l'évolution de ce graphe dans le temps (à partir des dates de publications).

Enfin, l'objectif final sera de mesurer des biais potentiels dans la structure de ce(s) graphe(s) temporel(s) liés par exemple au genre (homme/femme) ou à l'institution.

Rendu: Le rendu attendu est d'une part le graphe temporel, stocker dans un format approprié et le code permettant de faire les différentes analyses (exploration et visualisation). Les étudiants devront également réaliser un notebook jupyter de démonstration pour la partie exploration du graphe construit. **Le groupe de travail utilisera git pour mettre le code en commun**