Université de Sherbrooke

IFT 599 / IFT 799 - Science de données

TP3 : Exploration du web

Automne 2025

Enseignant

	Courriel	Local	Téléphone
Jordan F. Masakuna	jordan.felicien.masakuna@usherbrooke.ca	D6-0013	+1 873 662-8960

Faculté des Sciences,
Département d'informatique

26 novembre 2025

Sommaire

Dans le cadre de ce travail pratique (TP), un exemple de code consacré au classement de pages web ainsi qu'à l'analyse de réseaux complexes, accompagné d'un jeu de données, est mis à la disposition des étudiantes et étudiants. L'objectif est d'utiliser ce code et ce jeu de données, en suivant les consignes fournies, afin d'illustrer et de mettre en valeur les concepts liés à l'exploration du web présentés dans le Thème 7 du cours.

Contents

1 Objectifs pédagogiques

L'objectif de ce projet est de conduire une analyse expérimentale sur le classement de pages web (première partie) et sur l'étude des réseaux complexes (deuxième partie).

1.1 Classement de pages web

Dans cette première partie, les personnes étudiantes exploreront le fonctionnement des algorithmes PageRank et HITS et leur rôle dans le classement de l'information sur le web. Elles devront construire un graphe orienté à partir de pages web et de leurs hyperliens, puis appliquer ces deux algorithmes afin d'identifier les K pages les plus pertinentes (e.g., $K = 10$). L'analyse inclura une interprétation des résultats. Les personnes étudiantes seront sensibilisées aux contraintes pratiques du crawl, telles que la profondeur d'exploration et les biais induits par la requête initiale.

1.2 Minimisation des coûts de propagande simulée

Cette deuxième partie du projet consiste à concevoir une stratégie d'influence simulée dans un réseau social. Les personnes étudiantes mettront en pratique les métriques de centralité et les techniques de détection de communautés, notamment l'algorithme de Louvain. Elles devront sélectionner un sous-ensemble optimal de nœuds-influenceurs afin de maximiser la propagation de l'information tout en minimisant le coût global. L'exercice inclura la formulation et la résolution d'un problème d'optimisation sous contraintes de coût et de couverture.

1.3 Compétences développées

- Manipulation de réseaux complexes avec `python-louvain` et `NetworkX`.
- Comparaison et interprétation de métriques de centralité.
- Simulation et optimisation sous contraintes.
- Analyse critique et communication scientifique dans un rapport structuré.

2 Énoncé

2.1 Classement de pages web

Il est question de comprendre comment un moteur de recherche peut hiérarchiser l'information en fonction de la structure des liens, tout en sensibilisant aux limites pratiques liées au crawl, à la profondeur d'exploration et aux biais induits par la requête initiale.

Il s'agit d'explorer le fonctionnement des algorithmes PageRank et HITS à travers la construction d'un graphe d'hyperliens issu d'une requête web. À partir d'un mot-clé ou

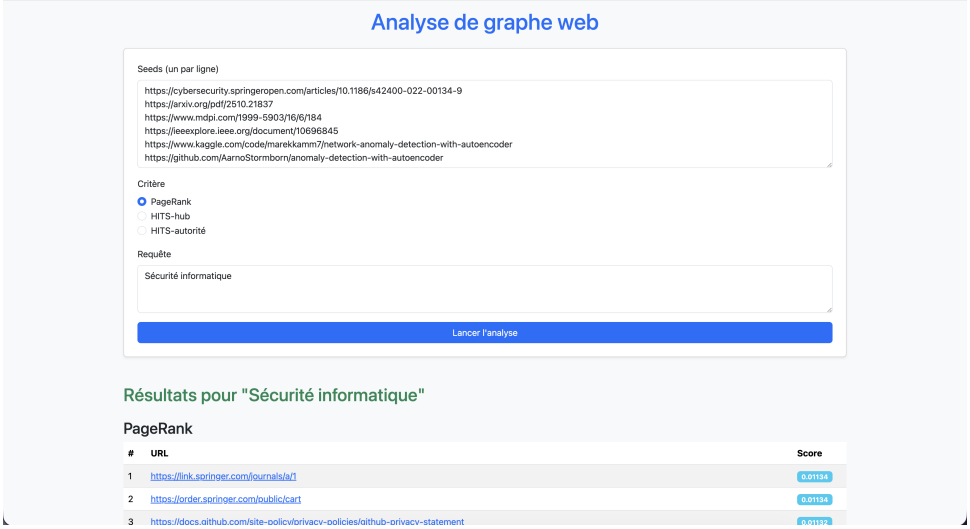

Figure 1: Explorateur du web

d'une expression de recherche, un *crawler* devra collecter un ensemble de pages pertinentes, extraire les liens sortants de chacune, puis représenter ces relations sous forme de graphe orienté. Sur ce graphe, les personnes étudiantes appliqueront PageRank et HITS afin de déterminer l'importance relative des différentes pages.

Techniquement, un crawler a toujours besoin d'un ou plusieurs points de départ (les seeds) pour explorer le web.

Il y aura donc deux critères de comparaison :

- PageRank des pages
- Autorité (par HITS) des pages

Un code a été partagé (voir `web.zip`—il peut être lancé avec la commande `python app.py`, ce qui ouvrira une page web en local sur le port 5000), qui implémente déjà le critère PageRank (comme illustré dans la Figure 1). Chaque groupe est censé implémenter le deuxième critère et adapter l'interface graphique au besoin. Comme pour PageRank, l'algorithme HITS est aussi implémenté dans la librairie `NetworkX`.

2.2 Minimisation de coûts de propagande simulée

Une personnalité politique souhaite mener une campagne électorale : faute de budget pour contacter l'ensemble des N citoyen.ne.s, elle doit cibler quelques individus capables d'influencer leurs communautés. On cherche à simuler une stratégie d'influence dans un réseau de N citoyen.ne.s, modélisés comme des utilisateurs d'un réseau social.

Chaque contact direct entre la personnalité politique et un.e citoyen.ne coûte en moyenne C_0 CAD (e.g., 200 CAD). En revanche, un contact réalisé par un·e influenceur·euse coûte C_1 ($C_1 \ll C_0$) CAD par citoyen.ne (e.g., 20 CAD), tandis que chaque influenceur·euse approché.e directement par la personnalité politique entraîne un coût de C_2 CAD ($C_0 \ll C_2$, e.g., 1000 CAD). L'objectif est de minimiser le coût total tout en

maximisant la couverture des citoyen.ne.s atteint.e.s dans le réseau :

$$\text{Budget idéal} = C_0 \times N \text{ CAD}$$
$$\text{Couverture idéale} = 100\%$$
$$\text{Budget simulé} = \left(C_2 \times n + C_1 \times \left| \bigcup_{v \in S} R_v \right| \right) \text{ CAD}$$
$$\text{Couverture simulée} = 100 \times \frac{\left| \bigcup_{v \in S} R_v \right|}{N} \, \%$$

où S l'ensemble des influenceur.euse.s choisi.e.s par la personne politique, $n = |S|$ le nombre d'influenceur.euse.s, et R_v désigne les citoyen.ne.s effectivement influencé.e.s par le nœud-influenceur v.

Les communautés sont détectées à l'aide de l'algorithme de Louvain, avec un nœud-influenceur sélectionné par communauté. Dans chaque communauté, le choix d'un.e influenceur.se repose sur une métrique de centralité. Chaque individu influent possède $p\,\%$ de chances de propager son influence à une personne membre de sa communauté.

À noter que les communautés obtenues avec l'algorithme de Louvain ne sont pas nécessairement connexes. Si un nœud influenceur ne peut pas atteindre un autre nœud faute de chemin entre eux, ce dernier ne peut pas être influencé.

Un notebook nommé `budget.ipynb`, contenant quelques implémentations préliminaires, a été partagé. La tâche demandée consiste à compléter ce notebook en respectant les consignes données.

3 Évaluation

3.1 Classement de pages web

- Réalisez trois à cinq requêtes (avec différents seeds pour chaque requête) et comparez, pour chacune, les dix premières pages obtenues selon chaque critère (en procédant à une lecture superficielle ou en survol des contenus des top K pages obtenues).
- Selon vous, quel critère permet d'identifier les pages les plus pertinentes ?
- Quel critère exige davantage de temps de calcul ? Vous pouvez utiliser la librairie `time` pour mesurer la durée d'exécution d'un bloc de code.
- Quel critère se révèle le plus sensible au choix des seeds ?

3.2 Minimisation de coûts de propagande simulée

À partir du réseau fourni et de la stratégie décrite :

1. Simulez la propagation de l'influence en sélectionnant un nœud-influenceur par communauté (détectée par l'algorithme de Louvain).

2. Pour chaque communauté, un nœud-influenceur est déterminé :
   - selon la centralité de proximité ;
   - selon la centralité de degré ;
   - selon la centralité d'intermédiarité (betweenness) ;
   - de manière aléatoire.
3. Pour chaque communauté et pour chaque méthode de détermination des nœuds influenceurs, calculez :
   - le budget simulé,
   - la couverture simulée,
   - le temps total de propagation de l'information (en supposant qu'un message prend η secondes pour passer d'un nœud à un autre, e.g., $\eta = 2$ secondes). Un nœud transmet l'information de manière séquentielle, un.e voisin.e après l'autre. Si un nœud influenceur v possède n_{vj} voisin.e.s de niveau j (où la distance entre le nœud et chacun.e. de ces voisin.e.s est j), alors le temps de propagation de l'information vers ces n_{vj} voisin.e.s est donné par :

$$T_{vj} = \eta \times n_{vj} \times j \text{ secondes.}$$

   Ainsi, le temps de propagation associé au nœud v est :

$$T_v = \sum_j T_{vj} \text{ secondes.}$$

   et le temps total de propagation sur l'ensemble du graphe est :

$$T = \sum_{v \in S} T_v \text{ secondes.}$$

   Par exemple, si un nœud v possède deux voisin.e.s direct.e.s (niveau 1) et trois voisin.e.s de niveau 2, le temps de propagation associé à ce nœud est (avec $\eta = 2$ secondes) :

$$T_v = 2 \times (2 \times 1 + 3 \times 2) = 16 \text{ secondes.}$$

4. Comparez les différents résultats obtenus selon l'approche de détermination des nœuds influenceurs, en termes de budget simulé, de couverture simulée, de temps total de propagation et de temps d'exécution. Proposez des visualisations pertinentes afin de faciliter la comparaison des résultats, notamment en termes de couverture perdue et de montant d'argent épargné par approche.

4 Travail à réaliser

4.1 Classement de pages web (40%)

Votre mission est de compléter le code fourni en implémentant la partie manquante : le calcul de l'autorité (HITS) pour chaque page du graphe. Vous pouvez enrichir le projet en ajoutant des fonctionnalités supplémentaires, telles que le choix du facteur d'amortissement α (pour PageRank) ou toute autre amélioration pertinente. Vous pouvez aussi optimiser le service de *crawler*.

4.2 Minimisation des coûts de propagande (40%)

Étant donné un réseau complexe, votre mission est de simuler le budget dont une personne politique pourrait disposer ainsi que la couverture atteinte. La tâche consiste à :

1. Appliquer l'algorithme de Louvain afin de déterminer le nombre de communautés ainsi que la composition de chaque communauté.
2. Sélectionner un.e influenceu.r.se par communauté, défini.e comme le nœud maximisant une métrique de centralité (i.e., la centralité de proximité, de degré ou d'intermédiarité) ou un nœud sélectionné aléatoirement.
3. Simuler la propagation de l'influence en supposant que chaque influenceu.r.se a $p\,\%$ de chances de convaincre une personne membre de sa communauté ($\frac{p}{100} \in [0.8, 1]$ aléatoirement et uniformément choisi). En d'autres termes, pour chaque membre de la communauté influencée par le nœud v, on effectue un tirage aléatoire—comparable au lancer d'une pièce biaisée—afin de déterminer si ce membre est convaincu ou non. En cas de succès (avec une probabilité de $\frac{p}{100}$), le membre rejoint l'ensemble R_v, i.e., l'ensemble des individus effectivement influencés par v.
4. Déterminez le budget et la couverture idéaux.
5. Calculez le budget et la couverture simulés.
6. Visualisez et comparez les résultats obtenus.

4.3 Présentation des résultats (20%)

Dans votre rapport, vous devrez présenter des commentaires détaillés sur les résultats obtenus. Si vous recourez à des ressources en ligne, il est impératif d'en citer les sources. L'absence de citation sera considérée comme un acte de plagiat et pourra entraîner l'attribution d'une note de zéro, en plus d'exposer l'étudiant ou l'étudiante à des mesures disciplinaires. Il est par ailleurs fortement déconseillé d'utiliser des ressources Internet pour la section consacrée à l'analyse des résultats.

5 Jeu de données et programmation

5.1 Jeu de données

YouTube est un site de partage de vidéos qui intègre un réseau social (`https://snap.stanford.edu/data/com-Youtube.html`). Dans ce réseau, les utilisateurs peuvent nouer des relations d'amitié et créer des groupes auxquels d'autres peuvent se joindre. Nous considérons ces groupes définis par les utilisateurs comme des communautés de référence. Le jeu de données (qui constitue un sous-graphe du réseau social YouTube) est inclus dans l'énoncé du TP et porte le nom `youtube.graph.edgelist`.

5.2 Programmation

Vous devrez utiliser `Python` comme langage de programmation. Les bibliothèques nécessaires pour ce TP ont été installées et utilisées lors de nos séances pratiques. Il est impératif de citer explicitement vos sources lorsque vous vous inspirez d'analyses réalisées par d'autres auteurs. Toute omission de citation sera considérée comme un acte de plagiat, pouvant entraîner l'attribution d'une note de zéro ainsi que des sanctions disciplinaires. Les citations peuvent être intégrées directement dans vos programmes sous forme de commentaires, ou regroupées dans une section dédiée de votre rapport, sous forme de liste de références.

6 Remise du TP

- Le TP doit être fait en équipe de deux à trois personnes.
- Indiquez les noms et Cips (ou matricules) des membres du groupe dans chacun des fichiers que vous soumettez.
- Ne téléchargez pas le jeux de données depuis les liens mentionnés ci-dessus. Utilisez plutôt la version incluse dans le fichier zippé du TP.
- La date de remise du TP est le mardi 09 décembre 2025 23h59'. Afin de garantir l'équité entre toutes et tous, aucun TP ne sera accepté après la date limite fixée.
- Soignez votre rapport, une pénalité de 5% pourrait être appliquée pour un rapport mal rédigé;
- Les fichiers à soumettre sont le rapport (en pdf) et l'ensemble de vos programmes. Ne pas soumettre les données!
- La remise doit être faite par Turnin : http://turnin.dinf.usherbrooke.ca