UNIVERSITÉ DE SHERBROOKE

DÉPARTEMENT D'INFORMATIQUE

**IFT 599/799 : Sciences de données**

# TP #4 — Automne 2021

# Recommandation par filtrage collaboratif

---

Ce travail vise à développer une application du filtrage collaboratif pour construire un système de recommandation. Les données à utiliser sont celles fournies par le MovieLens 100K. Ces données sont disponibles dans le répartoire public du cours. Le but de ce TP est de se familliariser avec une méthode de recommandation de base et la méthode de cross-validation pour évaluer la performance d'un système d'apprentissage.

La date pour la remise de ce TP est le vendredi 10 décembre 2021. Votre remise doit comprendre un rapport et les programmes que vous aurez développés pour ce TP, de même que des données modifiées ou nouvelles, s'il y a lieu. La remise doit se faire par la **turnin** sur https ://turnin.dinf.usherbrooke.ca/.

---

**Sujet : Recommandation par filtrage collaboratif**

Un répertoire de données est fourni dans le répertoire public du cours (Travaux/TP4/MovieLens). Les données à utiliser sont les u1.base, u1.test, ..., u5.base et u5.test. Les fichers .base contiennent les données d'apprentissage, alors que les fichiers .test contiennent les données pour tester les systèmes construits. *Pour ceux et celles ayant suivi le cours des techniques d'apprentissage (IFT603), ces 5 groupes de données .base et .test permettent de réaliser la validation croisée quintuple.* Par soucis de complétitude, tous les fichiers de cet ensemble de données sont inclus dans le répertoire, bien que plusieurs ne soient pas utiles pour ce TP. Vous devez lire le fichier "readme" pour connaître plus de précisions sur ces données.

Les principales tâches du TP sont les suivantes :

- Implantez un système de recommandation en se basant sur la méthode par voisinage, e.g. la méthode présentée dans la page 41 des diapositives du thème 6 ; Pour le calcul de la similitude entre les usagers, utilisez le coefficient de Pearson (page 52).
- Testez votre système sur les 5 groupes de données .base et .test. Choisissez l'une des deux mesures d'évaluation présentées à la page 37 des diapositives du thème 6. Donnez les résultats individuels (pour chaque ensemble de données de test) et la performance moyenne des 5 tests.

Pour ce TP, votre rapport doit être bref. Une page suffit. N'oubliez pas de remettre aussi vos données si elles sont modifiées des données originales.