Module 7 - Exemples d'applications des sciences des données

Exercices - Consignes

Cette série de questions à pour objectif de présenter les différents domaines d'applications de la science des données. Certaines questions nécessitent une exploration au delà des informations fournies dans le texte de référence du module 7. Dans ces cas, une recherche, via internet par exemple, est suggérée.

Exercices 1 : réseaux sociaux

L'objectif de cet exercice est de vous permettre de manipuler des graphes et d'appliquer des concepts liés aux graphes comme la centralité et le degré. Le texte de référence du module et ces exercices constituent un guide de démarrage pour vous dans ce domaine qui connaît une évolution importante.

Soit l'ensemble de données de Wikipédia (que vous pouvez télécharger à partir de ce lien `https://snap.stanford.edu/data/wiki-Vote.html`). Cet ensemble de données représente les données de vote des utilisateurs sur d'autres utilisateurs pour devenir des administrateurs. Prenons le fichier "Wiki-Vote.txt" (ce fichier est fourni avec cet exercice, donc vous n'avez pas besoin d'aller télécharger les données). Nous vous demandons de :

- calculer le nombre de noeuds (les votants) dans le graphe,
- calculer le nombre d'arcs (un arc donne une idée de qui vote sur qui) dans le graphe,
- calculer la transitivité du graphe,
- calculer la longueur moyenne des chemins entre deux noeuds dans le graphe,
- calculer la centralité moyenne normalisée des noeuds dans le graphe,
- identifier et visualiser les noeuds ayant les plus grandes valeurs de degré entrant. Ces noeuds représentent les utilisateurs qui ont reçu plus de votes.

Exercices 2 : intelligence d'affaires

L'objectif de cet exercice est de vous familiariser avec les données transactionnelles et les différentes méthodes de segmentation de clients. Plus spécifiquement, dans cet exercice, nous allons étudier une méthode qui s'appelle RFM (Récence, Fréquence, Montant, en anglais : *recency*, *frequency*, *monetary*). C'est une méthode très utilisée pour analyser la valeur des clients. La méthode RFM signifie [1] :

- Récence : date du dernier achat ou dernier contact client.
- Fréquence : fréquence des achats sur une période de référence donnée.
- Montant : somme des achats cumulés sur cette période.

Notez bien que cette méthode est descriptive, c'est-à-dire, elle ne permet pas de faire des prévisions, mais elle permet de bien segmenter les clients ce qui est très important dans le domaine du commerce [2].

Soit l'ensemble de données du e-commerce `data.csv` (disponible pour téléchargement sur le site du cours). Cet ensemble de données représente les données d'achats de clients dans un magasin sur une période allant du janvier 2010 jusqu'à décembre 2011. Nous vous demandons d'analyser ces données et segmenter les clients selon les trois critères : Récence, Fréquence, et Montant.

1. Définition tirée du Wikipédia
2. Cet exemple est également disponible sur le site de Kaggle au `https://www.kaggle.com/`