Initial Project Proposal: Detailed analysis and/or modeling of a network dataset

Notre groupe se compose de Damien BABET, Claudia DELGADO et Gabriele RANIERI.

LES DONNÉES

Les données sont disponibles sur ce site : http://konect.uni-koblenz.de/networks/bookcrossing_rating. Il s'agit du réseau bipartite de classification des livres de la communauté BookCrossing. Les notes sont des nombres entiers entre un et dix, où dix représente le meilleur score.

BUT DU PROJET

On se met dans la peau d'Amazon, et on essaye de segmenter la clientèle : faire de la détection de communautés (localisation des utilisateur/ genre des livres)

Question(s) pouvant être soulevée(s) par le détenteur des données (Amazon) :

- Quels sont les préférences en termes de littérature de nos utilisateurs ?
- Quels sont les utilisateurs qui ne contribuent pas à la base de données (i.e. qui ne notent aucun livre) ?
- Comment améliorer la distribution de nos livres ? Doit-on privilégier un genre ? Cette spécification a-t-elle un sens en termes de pays/régions où on parle la même langue ?
- Quels livres proposés aux clients en fonction de son segment de clientèle ? (dans une logique l'algorithme de prédiction)

Question(s) traitée(s) dans notre projet :

- A quoi ressemble l'espace de la littérature ?
- A quoi ressemble l'espace des lecteurs ?
- Les livres mieux notés se rapprochent-ils par genre, par langue, par année de parution ou par d'autres critères plus subtils comme le style, etc. ?
- Il y a des communautés de lecteurs spécifique par région/pays/langue parlés dans certain genre de livre ?

MÉTHODES MISES EN PLACE

Sous échantillon de travail :

Sur quelle base de travail doit-on travailler ?

- Tous les livres qui ont été noté ou uniquement les livres notés présents dans la base ?
- Toutes les notes ou uniquement les notes comprises entre 1 et 10 ?
- Les livres notés uniquement par des utilisateurs présents dans la base ?

Approches possibles :

Utilisateurs :

Au vus des données seule la localisation peut être implémenté sous forme de réseau. En effet les il y a 32% utilisateurs différents dans la base qui ont notés au moins un livre. Mais seul 60% ont renseigné leur âge alors que 100% ont renseigné leur lieu de résidence.

Livres :

Nous souhaitons enrichir la base de données grâce à différents attributs en les scrappant à partir d'internet (nombre de pages, style, genre, langue, année de parution)

ANNEXE :

		Remarque
Utilisateur (ils ont tous une localisation)	140 291	Un grand nombre d'utilisateurs sont présents dans la base mais n'ont noté aucun livre
Utilisateur avec âge	84 679	
Utilisateur ayant noté au moins un livre	44 778	
Notes (0 et 10)	493 813	Que faire des notes égales à 0 : Les garder ? ou les supprimer ? En effet on peut se demander si elles nous renseignent (ou non) sur le fait que l'utilisateur n'a pas du tout apprécié ce livre.
Notes égale à 0	317 794	
Notes entre 1 et 5 (inclus)	29 812	
Notes entre 6 et 10 (inclus)	146 207	
Livre	115 253	Certains ISBN ne sont pas bien renseignés
Livre ayant reçu au moins une note	204 680	Il y a des livres qui ont été noté qui ne sont pas dans la base de données
Livre ayant reçu au moins une note et n'étant pas présents dans la base initiale	67 665	

Tableau récapitulant les caractéristiques des données