Mini projet : K-moyennes pour le clustering de films

Nicolas Pécheux

Octobre 2014

1 Contexte

Vous êtes employés par un loueur de film qui aimerait réorganiser ses étalages. Il aimerait pouvoir ranger ses films dans différents rayons, de manière à grouper les films les plus similaires ensemble (en effet, son idée marketing est que lorsqu'un client choisit un film, son regard doit tomber immédiatement sur d'autres films qui lui donnent envie et qu'il voudra donc emprunter également). Ce commerçant hésite à investir dans 10, 50 ou encore 100 étagères pour classer ses films et aimerait avoir votre avis. De plus il dispose d'une base de donnné d'avis recueuillis auprès des ses clients et sait qu'un expert en Traitement Statistique de l'Information pourrait bien lui suggérer un classement optimal des films en différentes catégories.

2 Consignes

Ce TP noté (ou mini-projet) se déroule sur deux séances et est à rendre pour le **Dimanche 12 Octobre à 23h59**. Vous devez écrire un petit rapport (5 pages maximum) en expliquant en termes clairs les problèmes que vous vous êtes posés, ce que vous avez fait pour les résoudre et les résultats obtenus. Les figures et les tableaux de résultats simples sont les bienvenus s'ils servent vos propos. Vous pouvez travailler en binômes et rendre un rapport à deux. Vous devez envoyer votre rapport à `nicolas.pecheux@limsi.fr` *exclusivement* au format *Portable Document Format* (PDF) en utilisant la convention de nommage `nom1_nom2.pdf`. Votre code source ne sera pas évalué en tant que tel, mais doit être joint à l'envoi *impérativement* sous forme d'archive compressée `nom1_nom2.tar.gz`. Le non respect strict de ces consignes sera sévèrement pénalisé. Vous serez évalués sur la qualité du rapport, la méthodologie mise en place et les résultats obtenus.

3 Ce qu'il faut faire

Vous êtes libre de proposer les approches que vous voulez pour répondre aux questions du louer de film, tant que vous expliquez clairement ce que vous faites, pourquoi vous le faites, comment vous le faites et qu'est-ce que cela donne. Vous devez cependant au moins implémenter l'algorithme des K-moyennes vu en cours avec les données (films) du TP précédent. Vous pouvez utiliser la corrélation implémentée lors du TP précédent comme distance de similarité et la moyenne arithmétique de base pour (re)calculer les centroïdes. Parmi les pistes que vous pouvez explorer : Comment choisir les points initiaux ? Que se passe-t'il si on prend des points initiaux aléatoires et qu'on relance l'algorithme plusieurs fois ? Combien d'itérations faut-il faire ? Combien de clusters choisir ? Comment évaluer les résultats ?