Système de recommandation

Janvier 2017

1 Principe

L'objectif du TP est de réaliser un système de recommandation de films en utilisant le corpus MovieLens. [1] Ce corpus contient des jugements de films par différents utilisateurs ainsi que des informations sur les utilisateurs (âge, sexe, ...) et les films (réalisateur, genre, ...). Une version pré-traitée du corpus ne contenant que les informations que nous allons utiliser est disponible sur le site du cours. Son format est le suivant :

— un jugement par ligne ;

— trois champs par jugement (dans cet ordre) : l'utilisateur qui a donné la note, le film et la note ;

— les champs sont séparés par le caractère `|`.

2 Analyse du corpus

1. Donnez le nombre de jugements et le nombre d'utilisateurs présents dans le corpus.
2. Combien de films différents le corpus contient-il ? De quand date le film le plus récent ? le pus ancien ?
3. Représenter la distribution des notes. [2]
4. Caractériser le nombre de jugements par utilisateur en donnant la moyenne et la déviation standard du nombre de jugements par utilisateur ainsi que le plus grand et le plus petit nombre de jugements donné par un utilisateur.

3 Réalisation d'un système de recommandation

Le système de recommandation que nous proposons de développer repose sur notre capacité à déterminer (automatiquement) la similarité entre deux films. Il suffira alors de recommander un film similaire à celui qu'a aimé/consulté un utilisateur.

Le calcul de similarité d'une paire de films donnée peut se faire de la manière suivante :

— on extrait l'ensemble des n utilisateurs ayant jugé au moins un des deux films ;

1. `http://www.grouplens.org/node/73`
2. La bibliothèque la plus utilisée pour tracer des graphiques en python est `matplotlib`.

— chaque film est représentés par un vecteur de $\mathbb{N}^n$ dont la i^e coordonnée correspond à la note du i^e utilisateur ; si l'utilisateur n'a pas jugé le film, la coordonnée est nulle ;

— la similarité entre les deux films est donnée par la corrélation entre ces deux vecteurs.

5. Pourquoi la corrélation peut-elle être utilisée comme mesure de similarité ? Quelle autre mesure de similarité pourrait-on utiliser ? Donnez un avantage de la mesure de similarité « corrélation ».
6. Écrivez le code permettant de calculer la similarité entre tous les films du corpus.
7. Quelle est la complexité du calcul de la similarité entre tous les films du corpus considéré. Combien de temps a pris ce calcul sur votre machine ?
8. Donner les 5 films les plus similaires à `Scream (1996)` et à `Stargate (1994)` en précisant la valeur de leur similarité.
9. Comment évaluer la qualité des résultats ?