

Sujet3 : Recommending music and the odioscrobber data set

Réalisé par : GRINA Badreddine+ Salah halkhoums+Haitam Bahmed

1. Data Set:

L'ensemble de données Audioscrobber est un ensemble de données rendues publiques que les lecteurs peuvent acquérir sur un site web.

L'ensemble de données se compose de trois parties principales.

Le fichier user_artist_data.txt est le fichier principal de l'ensemble de données qui enregistre environ 2420 identifications d'utilisateurs, d'artistes et le nombre de fois que l'utilisateur écoute des chansons d'artistes, dont 141 000 utilisateurs et 1,6 million d'artistes ;

Le fichier artist_data.txt enregistre l'identification de l'artiste et le nom correspondant ;

Le fichier artist_alias.txt enregistre l'identification de l'artiste et l'identification de l'alias correspondant.

2. Recommended algorithm :

Comme l'ensemble de données sélectionné n'enregistre que l'interaction entre l'utilisateur et la chanson, il n'y a pas d'autre information que le nom de l'artiste. Par conséquent, l'algorithme d'apprentissage recherché ne nécessite pas d'informations sur les attributs des utilisateurs et des artistes. Ces algorithmes sont souvent appelés "filtrage collaboratif".

Si l'on estime qu'ils peuvent avoir des préférences similaires en fonction du même âge des deux utilisateurs, on ne parle pas de filtrage collaboratif. Il s'agit plutôt d'un filtrage coordonné, basé sur le fait que deux utilisateurs ont joué de nombreuses chansons identiques pour juger qu'ils peuvent tous aimer une chanson.

L'algorithme utilisé dans cet article est mathématiquement appelé les moindres carrés itératifs, et l'utilisateur joue les données comme la matrice A. La valeur de l'élément de la j^{ième} colonne de la matrice i-row basse représente la musique de l'utilisateur i jouant l'artiste j. La matrice A est clairsemée, et la plupart des éléments sont 0.

L'algorithme décompose A en deux petites matrices X et Y, toutes deux $A=XY^T$, X pour la matrice des caractéristiques de l'utilisateur, et Y pour la matrice des caractéristiques de l'artiste.

Maintenant, s'il y a 5 auditeurs et 5 musiques, alors A est une matrice 5×5 .