Product Data Scientist – Exercice Technique

Profil utilisateur

Afin d'analyser le comportement de nos utilisateurs nous récupérons l'ensemble de leurs écoutes réalisées sur Deezer depuis 2009.

Dans le cadre de l'exercice, nous supposerons que nous stockons ces écoutes dans des fichiers plats, chacun de ces fichiers contient les écoutes réalisées un jour J :

- Chaque fichier contient ~50M de lignes.
- Nous avons ~16M d'utilisateurs actifs.
- Notre catalogue contient ~35M de tracks.

Le format de ces fichiers est le suivant : <user_id>|<country>|<artist_id>|<track_id>
Exemple de ligne :
5464754|FR|542|79965994

1. Nous souhaitons représenter chaque utilisateur par son profil d'écoutes. Ce profil doit être maintenu à jour régulièrement et de volume raisonnable afin d'être utilisé par le site en production.
    a. Faire un schéma représentant la chaîne de traitement que vous proposez.
    b. Ecrire dans le langage de votre choix le code correspondant.

2. Nous souhaitons proposer aux utilisateurs de pouvoir « follow » sur le site des utilisateurs ayant des goûts musicaux similaires.
   Nous supposons que nous disposons d'une matrice **M** tel que M(i,j) est une mesure de similarité entre l'utilisateur i et l'utilisateur j.
    a. Définir une mesure de similarité. Justifiez.
    b. Quelle est la taille maximale de la matrice ? Quelles sont ses propriétés ? Comment choisiriez-vous de la stocker ?
    c. Ecrire dans le langage de votre choix le code correspondant à une fonction getSimilarUsers prenant en paramètre un <user_id> et retournant ses 20 utilisateurs les plus similaires.