Quel contenu Netflix est disponible dans les différents pays ?

Graphs algorithms and matching - Epitech - 2020

Antoine La Mache
Floriane Roy

1. Introduction

L'ensemble de données se compose d'émissions de télévision et de films disponibles sur Netflix à partir de 2019. Il est collecté à partir de Flixable, un moteur de recherche tiers de Netflix.

En 2018, ils ont publié un rapport qui montre que le nombre d'émissions de télévision sur Netflix a presque triplé depuis 2010. Le nombre de films du service de streaming a diminué de plus de 2000 titres depuis 2010, tandis que son nombre d'émissions de télévision a presque triplé. Il sera intéressant d'explorer ce que toutes les autres informations peuvent être obtenues à partir du même ensemble de données.

Les données sont classifiées dans les colonnes suivantes : l'ID, le titre, le genre, le directeur, le casting, le pays, la date d'ajout, la date de réalisation, la durée, la description, le classement.

Il est alors possible de se questionner sur le contenu en fonction des différents pays.

2. Métrique et graphique de similarité

Afin d'étudier cet ensemble de données, nous avons construit une métrique représentant les compatibilités entre les points de données. Dans un premier temps nous avons identifié les colonnes contenant des trous et les duplicatas. Ensuite nous avons supprimé les variables que nous avons jugé non pertinentes dans le cadre de notre analyse. Certaines colonnes ont aussi été remplies pour combler les informations manquantes.
Le graphique se base donc principalement sur les points suivants pour chaque film ou série : id, titre, pays, année de réalisation, date de réalisation.

L'année de réalisation a été convertie au format date et heure et ensuite au format Epoch afin de récupérer le nombre de secondes écoulées et faciliter les calculs.
L'ensemble contient des données quantitatives et non quantitatives, pour cela il a fallu transformer ces données pour pouvoir les exploiter.
Le calcul des distances euclidiennes est ensuite construit grâce à la formule suivante :

$$d(M_1, M_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Avec d la distance entre des points M1 et M2, et x et y leurs coordonnés.
Dans notre cas, M1 et M2 représentent le pays et l'année de réalisation du film.
On construit ensuite à partir des différences un graphique non orienté afin d'éviter de devoir prendre en compte les différents sens des arêtes. Ensuite, les points en utilisant l'id des films.

3. Regroupement des données et DBSCAN

Dans la mesure où le nombre de clusters n'était pas connu à l'avance, le choix a été fait d'utiliser l'algorithme de regroupement DBSCAN plutôt que le traditionnel K-means. Une des difficultés d'une telle approche fut de choisir une distance minimale entre les points appropriés. De manière à avoir des clusters cohérents. De plus, la longue génération du graph en $O(n^2)$ fut dans un premier temps une difficulté avant de décider d'importer directement la fonction de calcul de la métrique dans l'algorithme de clustering afin de pouvoir travailler directement en O(nlogn).

4. Analyse des résultats

L'algorithme DBSCAN a donné un cluster principal, correspondant à une majorité de contenus américains et quelques contenus de bi-nationalité (mais toujours possédant une partie américaine dans la production). Ce même cluster correspond en majorité aux productions hollywoodiennes, les autres étant sorties du giron et considérées comme des outliers. Ce qui peut donner, par exemple dans un algorithme de proposition de contenu à une ségrégation des contenus n'ayant pas été produit par une personne morale ou physique américaine et/ou dont la sortie ne fait pas partie d'une période temporelle bien donnée (nous pouvons donc penser par exemple aux anciennes production d'avant guerre qui ne seront pas proposées aux utilisateurs).

5. Conclusion

L'algorithme choisi présente une complexité de type $O(n^2)$ pour la génération du graph, ce qui produit un temps de calcul long et pourrait être amélioré en utilisant des moyens de parallélisation plus avancés (multi-threading, calcul sur GPU, …).