Netflix : visualisation des données

Projet de Visualisation des données - M1 MAS Rennes 2

Yolan PERONNET, Paul LANCELIN

12 mars 2021

Introduction

Le 2 octobre 2006, Netflix lançait une compétition ouverte pour le meilleur algorithme de filtrage collaboratif pour prédire les notes des utilisateurs pour les films. Le 21 septembre 2009, après avoir mis plus de 40 000 équipes en concurrence durant 3 ans, le grand prix de 1 000 000 $ US a été remis à l'équipe Pragmatic Chaos de BellKor qui a battu le propre algorithme de Netflix. Afin de réaliser de tels algorithmes de prédictions, les équipes ont dû au préalable comprendre les données Netflix et ce, en passant notamment par une phase de visualisation de ces données. L'objectif de notre projet est donc de réaliser une application permettant une première visualisation des données concernant le contenu présent sur Netflix, pouvant ainsi aider des équipes souhaitant développer un nouvel algorithme de prédiction dans leur phase de compréhension du contenu présent sur la plateforme. Ils pourront grâce à elle tirer certains résultats qui leur permettront d'affiner l'orientation de leur algorithme. De plus, la sollicitation de cette plateforme ayant explosé depuis le premier confinement en 2020, cette application permettra aux utilisateurs les plus curieux d'appréhender l'ensemble du contenu que Netflix propose sur sa plateforme. Enfin, une section de l'application donnera la possibilité de confronter le contenu présent sur Netflix, Hulu, Amazon Prime et Disney+.

Ce document aura donc pour but de présenter ce qui a été réalisé au cours de ce projet de visualisation du contenu présent sur Netflix tout en pointant les différentes difficultés rencontrées. Dans un premier temps, nous présenterons les différentes tables auxquelles nous avons eu accès constituant ainsi notre base de données. Par la suite, seront abordés les principaux axes de visualisation que nous avons décidé de mettre en lumière. Enfin, une présentation de l'application RShiny et de son fonctionnement sera faite.

1. Présentation de la base de données

Trois tables récupérées sur kaggle nous ont permis de constituer notre base de données. Toutefois, nous aurions aimé augmenter cette base de 2 autres tables : une première permettant d'analyser les volumes de temps de streaming et le nombre d'utilisateurs et une seconde correspondant à la table fournie par Netflix pour sa compétition. Cependant, la première n'a pas été trouvée et l'analyse de la seconde aurait demandé un traitement préalable conséquent pour pouvoir être exploitée. Nous nous sommes donc restreint à l'analyse de 3 tables que nous vous présentons dans les paragraphes suivants.

1.1. L'ensemble du contenu présent sur Netflix au 16 janvier 2021 : netflix_titles

Cette première table est celle qui nous a permis de réaliser la majeure partie de notre projet. Elle est constituée de 11 variables et de 7787 individus, chaque individu correspondant à un film ou une série disponible sur Netflix. Pour les variables, nous avons le type du contenu (Movie ou TV Show), le titre du film, le ou les réalisateurs (sous forme d'une chaîne de caractère, les réalisateurs étant séparés par une virgule), les acteurs (même format que pour les réalisateurs), le ou les pays de production (même format que les réalistaurs), la date d'ajout du contenu sur la plateforme (format : "Month XX, XXXX"), l'année de sortie du contenu, la durée du film si film (format : "X min") ou le nombre de saison si série (format : "X Season(s)"), la caractérisation du contenu (Parental Guidance Suggested...), le ou les genres (même format que les réalisateurs) et la description du film.

Pour simplifier les futurs traitements, nous avons transformé la variable de date d'ajout au format date, nous avons décomposé la variable de "durée" en 2 variables (nombre de saisons si série et durée du film en minutes si film), et nous avons homogénéiser la dénomination des pays (par exemple West Germany devient Germany, Soviet Union devient Russian Federation etc.).

Lien de la table : https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

1.2. L'ensemble des films présents sur 4 plateformes de streaming en 2020 : movies_platforms

La seconde table a pour objectif de recenser tous les films présents sur les 4 principales plateforms de streaming : Netflix, Hulu, Prime Video et Disney+. Cette seconde table comporte 15 variables et 16 744 films. Les variables sont les suivantes : titre, année de sortie, âge recommandé, note IMDb, note "Rotten Tomatoes", 4 variables indicatrices pour les 4 plateformes (1 si le film est présent sur la plateforme, 0 sinon), le ou les directeurs, le ou les genres, le ou les pays de production, la ou les langues disponibles pour le film, et la durée.

Lien de la table : https://www.kaggle.com/ruchi798/movies-on-netflix-prime-video-hulu-and-disney

1.3. L'ensemble des séries présentes sur 4 plateformes de streaming en 2020 : tv-shows_platforms

Cette dernière table est exactement la même que la précedente, mais pour les séries (5 611) et diminuée des variables : le ou les directeurs, le ou les genres, le ou les pays de production, la ou les langues disponibles pour le film, et la durée.

Lien de la table : https://www.kaggle.com/ruchi798/tv-shows-on-netflix-prime-video-hulu-and-disney

2. Présentation des principaux axes de visualisation

2.1. Visualisation de la table netflix_titles

2.1.1. L'ajout du contenu Netflix en fonction du temps

Tout d'abord, nous avons voulu visualiser les caractéristiques temporelles de l'ajout de contenu sur Netflix. Pour se faire, nous avons visualisé le nombre de films et séries sur la plateforme en fonction des années, et nous avons visualisé également le nombre d'ajouts de séries et films sur la plateforme en fonction des mois de l'année (avec la possibilité de choisir un filtre sur le genre). L'objectif de cette dernière figure est de par exemple regarder à quel moment de l'année sont ajoutés les films de

genre romantiques. Enfin, nous avons visualisé par un diagramme en bâtons de la distribution des écarts “ajout sur Netflix - date de sortie” en années.

2.1.2. L’ajout du contenu Netlfix en fonction des pays de production

De plus, nous avons voulu visualiser quels étaient les pays ayant produit le plus de films/séries présents sur Netflix. Pour se faire nous avons réalisé des cartes (statiques ou interactives) donnant le nombre de films ou séries (au choix) sur Netflix produits dans chaque pays, pour un genre choisi. Pour continuer dans la caractérisation spatiale, nous avons également produit 2 autres figures (diagramme en barre des 10 pays ayant le plus produit pour un genre choisi et graphe “à bulle” des 10 pays ayant le plus produit).

2.1.3. La durée des films et le nombre de saisons par série

Ensuite, nous avons voulu visualiser la distribution de la durée des films en fonction de l’année de sortie et du genre du film. Pour se faire, nous avons d’abord réalisé un histogramme des durées avec les 2 filtres choisis, et nous avons également représenté l’évolution de la durée moyenne en fonction des années.

Nous avons également voulu visualiser le nombre de saisons par série et avons pour cela tracer un diagramme en bâtons avec (nombre de saisons en abscisse).

2.1.4. Les tops N (genres, directeurs, acteurs)

Enfin, les derniers traitements de la table netflix_titles consistaient en la visualisation de “top N”. Nous avons donc pour cela réaliser des diagrammes en barre des N genres les plus présents sur la plateforme (choix du N), des N réalisateurs ayant produit le plus de contenu présent sur la plateforme (choix du N, du pays et du genre), et des N acteurs ayant le plus joué dans des contenus présents sur la plateforme (choix du N, du pays et du genre).

2.2. Visualisation des tables tvshows_platforms et movies_platforms

2.2.1. Volume du contenu présent sur chaque plateforme en 2020

Tout d’abord, nous avons voulu visualiser le volume de contenu de chaque plateforme. Pour se faire, nous avons réalisé des diagrammes en barres avec le choix de visualiser le nombre de séries, le nombre de films ou les deux.

2.2.2. Ancienneté du contenu présent sur chaque plateforme

Ici, nous avons voulu visualiser l’ancienneté du présent sur chaque plateforme. Pour se faire, nous avons réalisé pour chaque plateforme un diagramme en bâtons avec le volume de contenu pour chaque année de sortie. Nous avons par ailleurs intégré la possibilité de choisir de la plage d’années des diagrammes en bâtons.

2.2.3. Les notes “Rotten Tomatoes” et IMDb

Dans cette partie, la visualisation de la distribution des notes “Rotten Tomatoes” et IMDb a été faite pour chaque plateforme à l’aide d’histogrammes.

Également, nous avons voulu vérifier la cohérence entre les deux types de notes auxquelles nous avions accès en réalisant une régression linéaire entre les 2 vecteurs de notes.

Enfin, pour chaque type de note, nous avons voulu regarder s'il existait un lien entre les années de sorties et les notes (les anciens films, sont-ils mieux ou moins bien notés ?). Pour se faire, nous avons donc réalisé 2 régressions linéaires.

2.2.4. Top 20 (réalisateurs, genres, pays) en fonction des plateformes

Ensuite, nous avons voulu visualiser, à l'image du 2.1.4, quels étaient pour chaque plateforme les 20 réalisateurs, genres et pays les plus présents. Pour cela, nous avons, pour chaque plateforme, construit une table réunissant ces trois top 20.

2.2.5. Sur quelle(s) plateforme(s) de streaming pouvais-je trouver ce film en 2020 ?

Pour finir, nous avons voulu créer un petit moteur de recherche prenant en entrée un titre de film ou de série et renvoyant la ou les plateforme(s) sur le(s)quelle(s) nous pouvons retrouver ce film. La possibilité de choisir certains filtres sera également intégré.

3. Présentation de l'application RShiny et de son fonctionnement

Vous pouvez retrouver l'application en cliquant ici ou avec l'url : https://paul-lancelin.shinyapps.io/Netflix/ .

L'application est structurée à l'aide du format dashboard avec un side panel comportant notre menu déroulant et d'un main panel. Le side menu est fixé pour plus de confort dans la transition entre les parties et peut aussi être caché. Notre menu est constitué de 4 grandes parties, une page d'accueil, une partie sur Netflix, une partie sur Netflix mise en parallèle avec les autres plateformes et un "moteur de recherche".

La page d'accueil contient une vidéo YouTube sur l'histoire de Netflix, elle est présente plus à titre de curiosité que d'intérêt pour la compréhension des données.

La partie Netflix Analysis se divise comme suit :

- Generality (cf 2.1.1)
- Netflix and countries (cf 2.1.2)
  - Plotting
  - Static Map
  - Interactive Map
  - Top (cf 2.1.4)
- Duration analysis (cf 2.1.1 et 2.1.3)
- Data (cf 1.1)

La partie Netflix vs Others se divise comme suit :

- Market Share (cf 2.2.1)
- Age of content (cf 2.2.2)
- Rating analysis (cf 2.2.3)
  - IMDb analysis
  - Rotten Tomatoes analysis
  - IMDb/RT linear model
  - Rating/Year linear model
- Top 20 (cf 2.2.4)

La dernière partie est Search Engine, l'utilisateur est amené à renseigner un nom de film ou de série qu'il souhaite regarder. L'input est fait pour suggérer en même temps que l'utilisateur tape le nom et en fonction des titres contenus dans les bases de données. La sortie est faite pour renseigner sur quelle plateforme le titre est disponible ou s'il n'est pas disponible. Deux autres facteurs sont modifiables, le type (film/série) et la période en années. (cf 2.2.5)

L'application s'organise en trois fichiers, ui, server et global. Le fichier global contient le traitement des données pour que les deux autres se concentrent sur la structure de l'application. Presque tous les graphes sont interactifs et certains réactifs à des inputs que l'utilisateur du site peut renseigner. Pour plus de confort d'utilisation, nous avons mis des boutons pour les graphiques où il y a plus d'un paramètre modifiable.
Le présent document est lui-même accessible depuis la page d'accueil et la carte statique est téléchargeable au format png à l'aide d'un bouton et le nom de fichier est réactif aux inputs renseignés par l'utilisateur.

Conlcusion

Nous avons apprécié réaliser ce projet en binôme, il nous a permis de mettre en commun nos connaissances et compétences. Le choix du sujet nous a permis de mettre en pratique la majeur partie de ce que nous appris en cours de visualisation de données et d'explorer les possibilités de création et de personnalisation d'application R Shiny.

Nous sommes globalement contents du rendu, mais remarquons néanmoins certaines pistes d'amélioration. De plus, nous aimerions fournir à l'utilisateur la possibilité de choisir entre différentes versions (langue : français/anglais, graphes : statiques/interactifs). La langue pour toucher un public plus large et le choix d'interaction sur les graphes pour améliorer l'expérience sur smartphone, le côté interactif peut s'avérer embêtant sur smartphone. De plus, nous faisons le constat que l'application semble relativement lourde au niveau du chargement. Nous avons essayé à l'aide de l'utilisation de fonctions de simplifier notre code, mais il reste sûrement d'autre aspect de R Shiny à comprendre pour optimiser la structure même de notre application (création de modules, etc).