GIRAUD Claire
GHESTEM Florence
GUISNEUF Virginie

Computer Science for Big Data

Amazon Prime et Netflix et enjeux sociétaux

I – Introduction et problématique

Les plateformes de vidéo à la demande connaissent un essor de plus en plus important en France mais aussi dans le monde, accéléré par la crise sanitaire. D'après le rapport de Statista du 11 novembre 2021, les plateformes de vidéos à la demande (VAD) comptaient en octobre 2021, 214 millions d'abonnés sur Netflix, 200 millions sur Amazon Prime et 118 millions sur Dinsey +. Le choix du contenu proposé et visionné par les abonnés de ces plateformes, a alors un impact social important.

Nous nous sommes demandées si ces plateformes proposent du contenu éthique ou engagé ? Pour trouver des éléments de réponse à cette question nous nous sommes intéressées à l'origine de production des films, à la diversité des réalisateurs et aux thématiques abordées par les programmes.

Cette problématique est d'autant plus intéressante que le catalogue évolue au cours du temps. Or, le catalogue de ces plateformes semble ne cesser d'augmenter. Nous avons alors cherché à étudier ces jeux de données de façon reproductible sur une base de données de plus en plus volumineuse dans la perspective d'étudier cette problématique à nouveau dans quelques années.

Dans un premier temps nous présenterons les données puis les outils utilisés pour répondre à notre problématique. Les résultats seront alors présenté d'un part ceux obtenus grâce au système de gestion MongoDB et d'autre spart ceux obtenus sur Spark. Ces résultats nous permettrons ensuite de conclure.

II – Choix et présentation des données

Pour répondre à cette problématique, nous avons voulu nous intéresser aux deux géants de la VAD : Netflix et Amazon Prime. Pour chacune de ces deux plateformes nous avons utilisé le catalogue USA comportant l'ensemble des films et séries disponibles. Ci-dessous les informations données pour chaque contenu et leur libellé :

- Le type (film ou séries) : *type*
- Le titre : *title*
- Le(s) réalisateur(s) : director
- Les principaux acteurs : *cast*
- Les pays de diffusion : *country*
- La date d'arrivée sur la plateforme : *date_added*
- La date de réalisation : *release_year*
- Le temps de film ou nombre de saison : *duration*
- Le genre (ex : comédie/drame…) : *listed_in*
- Une petite description : *description*

Les jeux de données *amazon.json* et *netflix.djon* utilisés n'ont pas reçu de traitement particulier dans leur contenu en amont des analyses. Ils sont disponibles sur le site Kaggle.

III - Choix des outils

Dans un premier temps pour réaliser des requêtes simples nous avons choisi d'utiliser le système de gestion MongoDB. MongoDB permet d'utiliser directement le format JSON et de manipuler aisément les données sans avoir à créer des tables. L'utilisation d'une base de données NoSQL facilite également la réplicabilité, facilitant la comparaison de nos deux plateformes. Par ailleurs, notre jeu de données ne présente pas de lien entre les variables, c'est pourquoi nous n'utilisons pas neo4J.

La fonction *aggregate* est utilisée dans ce projet afin de remplacer partiellement les algorithmes de map-reduce. Elle permet une facilité de lecture et d'utilisation pour des opérations assez simples comme c'est le cas dans la première partie de ce projet. Pour des plus grosses requêtes, le map reduce est nécessaire. En effet, les étapes d'agrégation ne peuvent pas consommer plus de 100Mo en mémoire (documentation mongodb).

Enfin, nous souhaitions réaliser des requêtes plus complexes et pouvoir étendre nos requêtes à un jeu de données éventuellement plus conséquent. Nous avions alors besoin d'un ou plusieurs systèmes de gestion de données capables d'effectuer des requêtes suivant le concept de map-reduce. D'une part, nous avons utilisé les fonctionnalités de map-reduce de MongoDB afin de les comparer à sa fonction *aggregate*. D'autre part, nous avons utilisé le système Spark afin de réaliser des requêtes en map-reduce sur ce qui est considéré comme l'avenir de la plateforme Big Data (aspect de reproductibilité).

IV - Résultats

1- Requêtes simples avec MongoDB

Pour commencer nous nous intéressons aux types et aux catégories de contenu sur les plateformes Netflix et Amazon Prime.

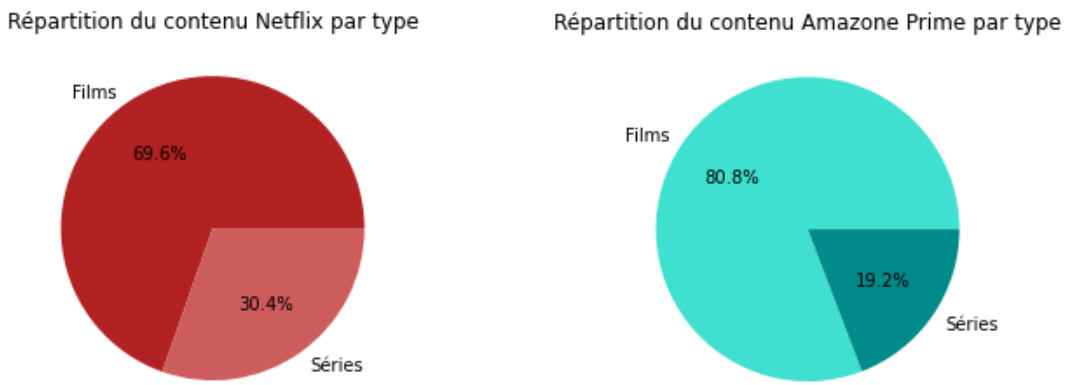

Figure 1 : Comparaison des types de contenu sur les deux plateformes.

Ce premier graphique (Figure 1) permet de connaître contenu que nous analysons ensuite. Amazon Prime diffuse une part plus importante de films que Netflix. A noter : Amazon propose également des clips vidéos classés comme 'films' ce qui explique la part importante de ce type de contenu.

GIRAUD Claire
GHESTEM Florence
GUISNEUF Virginie

Afin de savoir si les deux plateformes s'intéressent aux thématiques de sociétés (culture, causes sociales, environnementales, etc.) les catégories de films qui peuvent correspondre à ses problématiques sont recherchées.

Figure 2 : Comparaison par catégories de contenu sur les deux plateformes.

Sur la Figure 2 les deux plateformes ne catégorisent pas de la même façon leur contenu. Cependant on remarque que la catégorie "Autre" représente presque la même proportion (75%). Sur ce point les deux plateformes ne semblent pas se différencier. Un quart du contenu semble être centré vers des thèmes de société.
NB : Les shows de télé-réalités ont été exclus des docu-séries pour la plateforme Netflix.

Par la suite, nous nous sommes intéressés aux descriptions des films et plus précisément aux mots qu'ils contiennent. Nous avons choisi de compter l'occurrence de mots relatifs aux problèmes environnementaux (climate change, ecology, environnementale et global warming) et aux droits humains (harassement, police & violence, racism, women & rights). Pour cela l'opérateur $regex est utilisé pour chercher une expression régulière dans des chaînes de caractères.

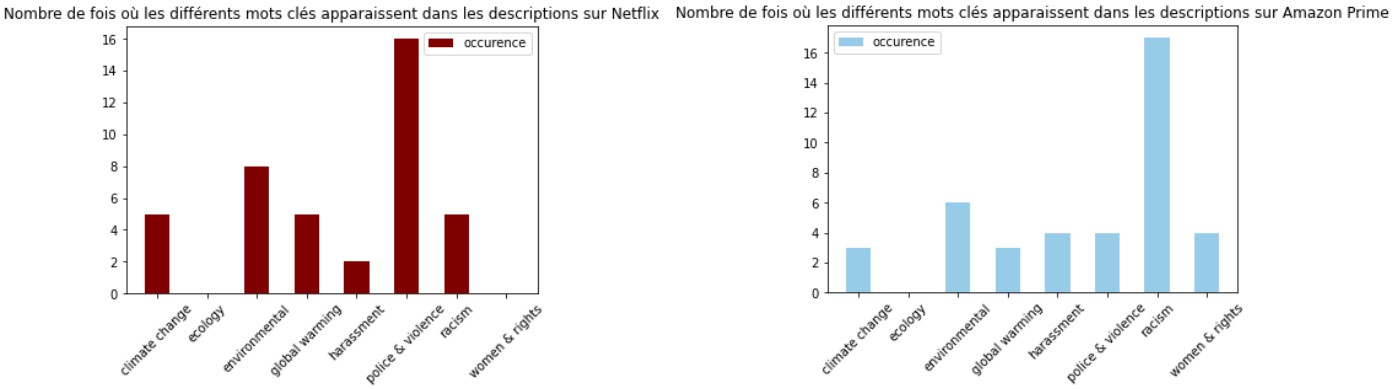

Figure 3 : Occurrence des mots clés trouvés dans les descriptions des différents films et séries sur les deux plateformes de streaming.

Sur les graphiques de la Figure 3, la même tendance est observée pour les deux plateformes. Le mot le plus récurrent dans les descriptions étant '"racism". Les thématiques environnementales semblent moins mentionnées que celles sociales.

Les étapes suivantes ont nécessité l'utilisation de la fonction aggregate. Un film pouvant être réalisé dans plusieurs pays et/ou par plusieurs réalisateurs, nous avons utilisé l'opérateur $split suivi d'un $unwind afin d'obtenir le nombre de films par pays puis par réalisateurs.

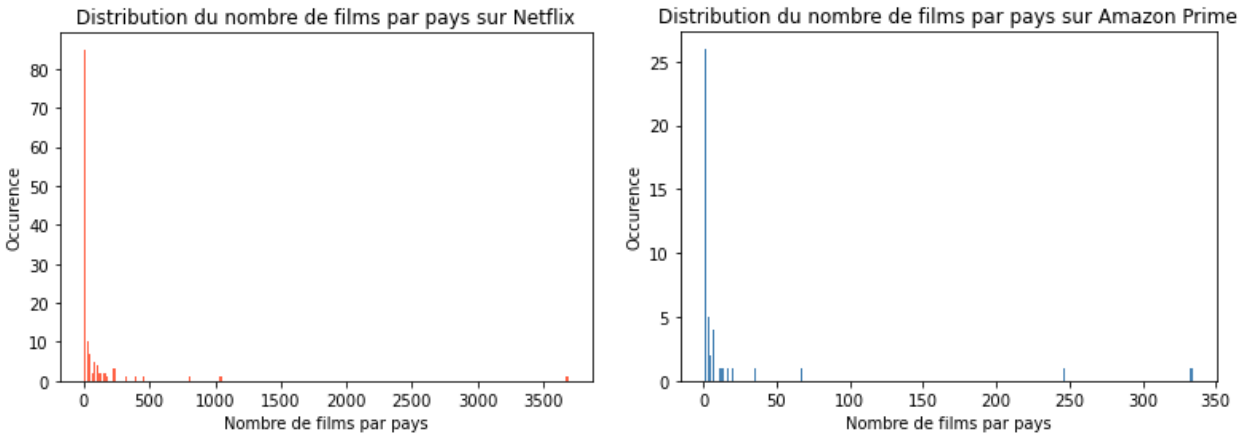

Figure 4 : Distribution du nombre de films par pays sur les deux plateformes.

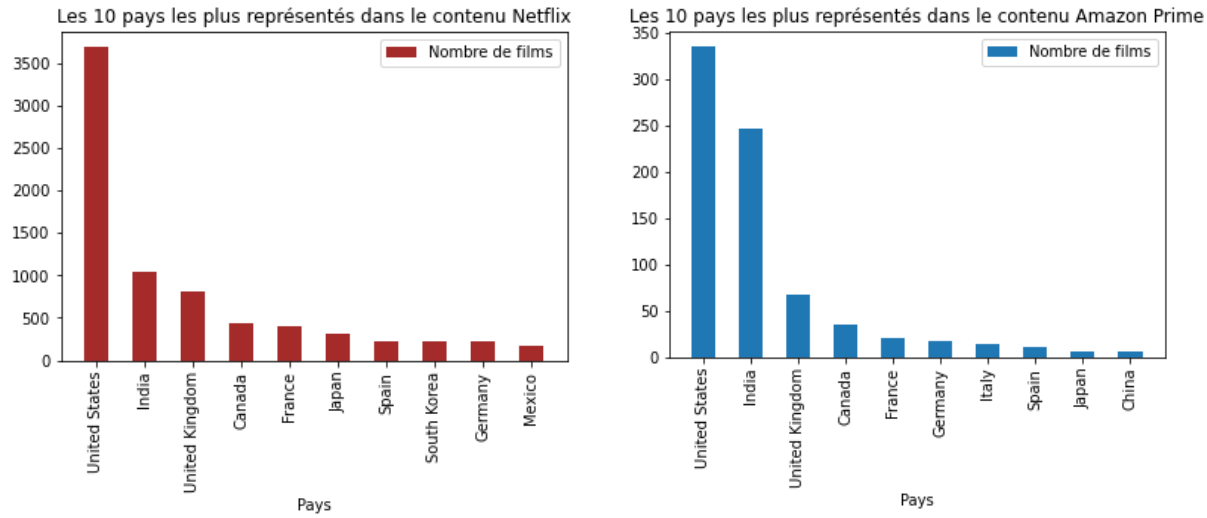

Figure 4 bis : Top 10 des pays les plus représentés sur les deux plateformes.

Sur les Figure 4 et 4bis la même répartition est observée pour les deux plateformes. Amazon Prime diffuse cependant plus de films Indien au détriment des autres pays. Bien que 126 pays soient représentés sur Netflix, la majorité ont moins de 50 films sur les 8 807 présents dans le catalogue. Ceci peut être expliqué par le fait que le jeu de données soit le catalogue de Netflix USA.

NB : Pour la plateforme Amazon Prime, la grande majorité (8996) des films n'ont pas de champ "country", ce qui biaise l'analyse.

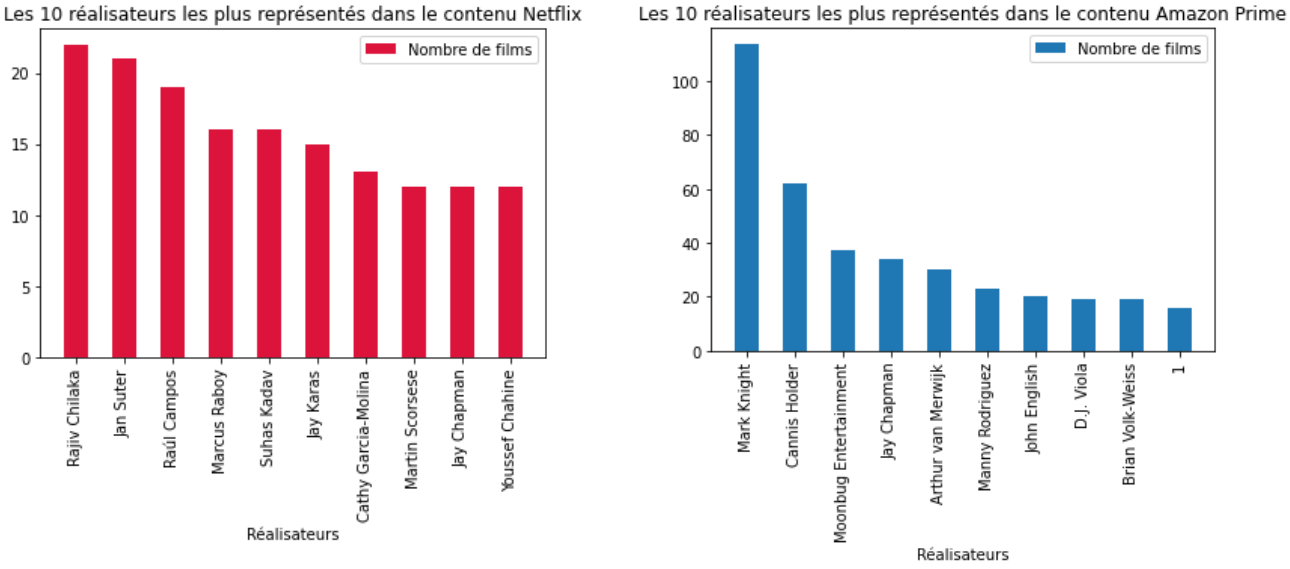

Figure 5 : Distribution du nombre de films par sur les deux plateformes.

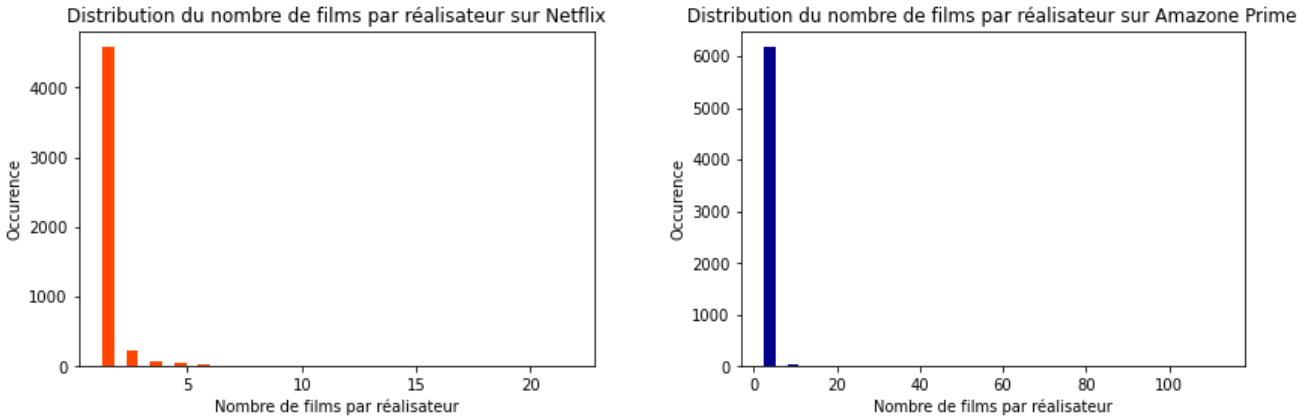

Figure 5 bis : Top 10 des réalisateurs les plus représentés sur les deux plateformes.

Au regard des Figure 5 et 5 bis, encore une fois, peu de différences sont observées entre Netflix et Amazon Prime. La plupart des réalisateurs ont réalisé moins de 5 films par plateforme. Il semble y avoir une bonne diversité.

2- Map-Reduce avec Spark

Nous avons souhaité étudier l'évolution du contenu culturel des plateformes. Afin de comparer les tendances en termes de réalisation et de diffusion, le catalogue a été étudié en fonction des dates de réalisation et de publication sur les plateformes. Tout d'abord, nous avons observé la catégorie de films indépendants.

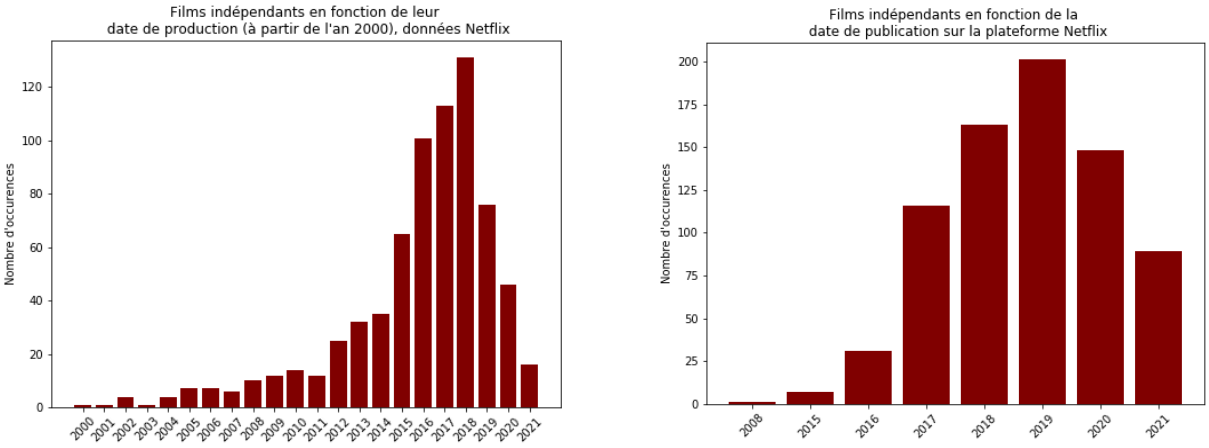

Figure 6 : Distribution du nombre de films indépendants par an (a : date de production, b : date de publication) disponibles sur Netflix.

Nous pouvons voir sur ces figures que la plateforme Netflix (Figure 6) a diffusé le plus de films indépendants en 2019 avec l'enrichissement du catalogue de 200 films environ. De plus, Netflix ne diffusait aucun film indépendant avant 2008. En parallèle, la période de production de films indépendants favorisée sur la plateforme se situe entre les années 2016 et 2018. Ce décalage au niveau des pics de production est de diffusion peut être expliqué par le délai entre la réalisation d'un film indépendant et sa promotion voire son succès. D'autant plus que dans le cas des films indépendants, le budget de production et de promotion est nettement plus faible. Cette analyse n'a pas pu être confirmée dans le cas de la plateforme Amazon prime car elle ne possède pas de films indépendants ou d'un genre assimilé.

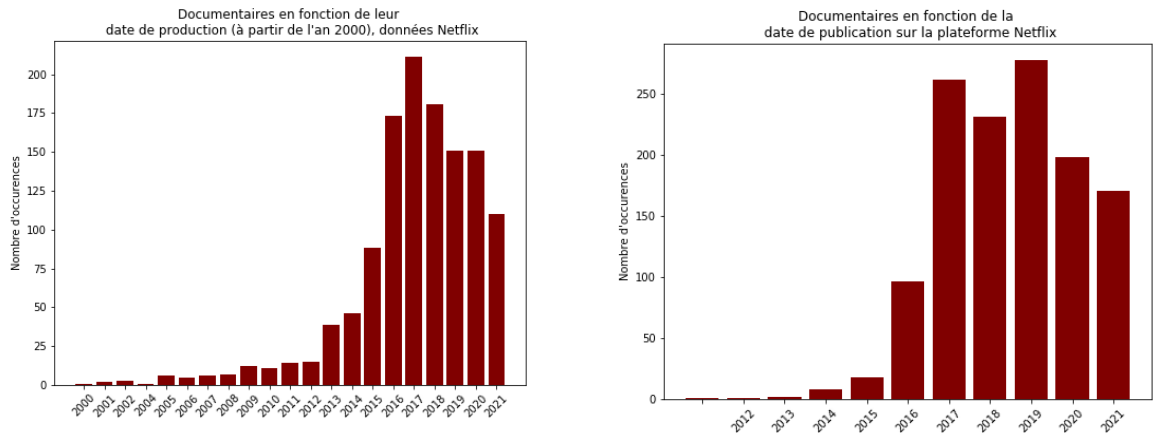

Figure 7 : Distribution du nombre de documentaires (films et séries) par an (a : date de production, b : date de publication) disponibles sur Netflix.

Nous avons alors choisi d'étudier une catégorie présente sur les deux plateformes : les documentaires. La tendance observée ci-dessus est la même pour la plateforme Netflix (Figure 7). En effet, les documentaires diffusés sont plutôt réalisés entre 2016 et 2018 et diffusés plutôt en 2019.

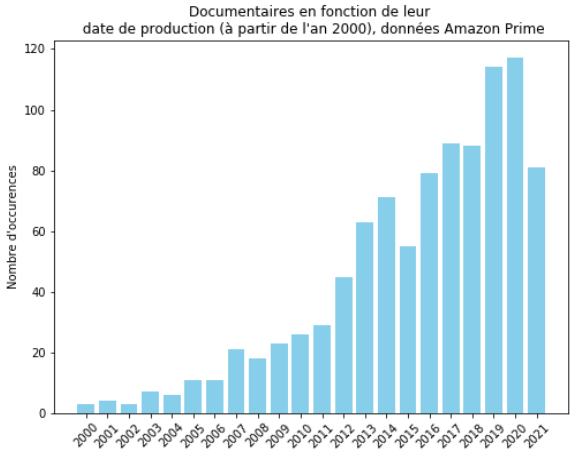

Figure 8 : Distribution du nombre de documentaires (films et séries) par an (date de production) disponibles sur Amazon Prime.

Concernant les données d'Amazon prime (Figure 8), nous pouvons aussi constater que les documentaires actuellement présents sur la plateforme ont été réalisés plus récemment (entre 2018 et 2020). Cependant, nous ne pouvons pas comparer les données des dates de diffusion sur Amazon prime car ce sont en majorité des données manquantes. Les documentaires disponibles ont une date de production récente. L'absence des données de leur diffusion peut être liée à un renouvellement trop récent de la base de données qui ne s'est peut-être pas actualisée.

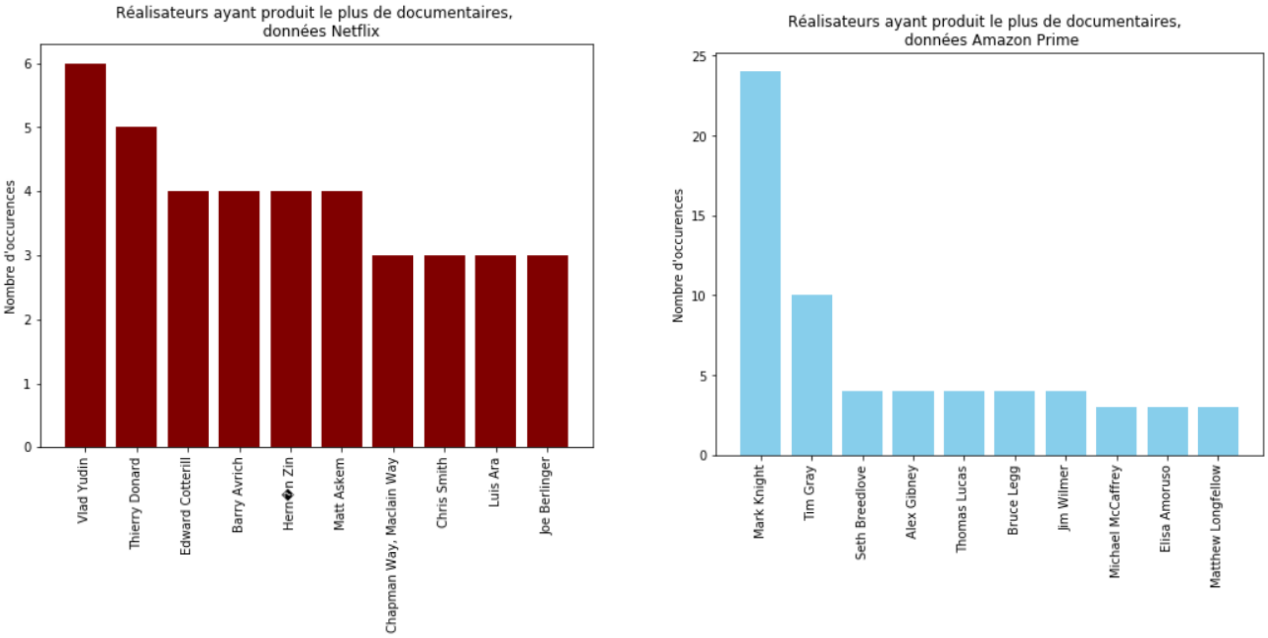

Figure 9 : Distribution du nombre de documentaires (films et séries) par réalisateur disponibles sur Amazon Prime et Netflix

Afin de voir la diversité des points de vue des documentaires diffusés, nous nous sommes intéressés aux réalisateurs de documentaires favorisés sur ces plateformes. Ainsi, la Figure 9 présente les 10 réalisateurs ayant produit le plus de documentaires diffusés sur les plateformes Netflix et Amazon prime. Du côté de Netflix, les réalisateurs sont représentés de façon assez équivalente, avec environ 4 films par réalisateur. En revanche, du côté d'Amazon prime, le réalisateur Mark Knight se distingue nettement des autres. En effet, il cumule 5 fois plus de documentaires publiés que ses collègues. Après quelques recherches, Mark Knight est en fait un professionnel ayant 25 ans de carrière en vidéo et photographie (Wilmer, s. d.). De même, Tim Gray, le deuxième réalisateur le plus influent sur Amazon prime, a produit au cours de sa carrière 28 reportages documentaires (World WarII Foundation, 2020). Nous pouvons ainsi en déduire que les quelques documentaires diffusés sur Amazon prime sont issus de réalisateurs reconnus. La tendance est nettement différente sur Netflix avec comme premier réalisateur Vlad Yudin, un réalisateur russe assez jeune.

Conclusion

Les diffusions sur les deux plateformes sont proches sur de nombreux aspects. En effet un quart de leur contenu est consacré à des thèmes sociétaux, l'occurrence des mots relatifs aux droits humains et aux problèmes environnementaux est similaire. Aussi, les deux plateformes diffusent de nombreux réalisateurs de manière équilibrée. Les pays représentés sont très majoritairement les Etat-Unis puisqu'il s'agit du catalogue USA. L'Inde est cependant très représentée sur Amazon Prime. Les autres pays sont faiblement représentés mais de manière plutôt équitable surtout sur Netflix. Au niveau des documentaires Amazon Prime diffuse plus ceux réalisés par un réalisateur connu, Mark Knight, alors que Netflix offre plus de diversité mais avec des réalisateurs moins renommés. Ainsi les deux plateformes semblent faire l'effort de proposer du contenu éthique et engagé en proposant des contenus diversifiés et traitant des enjeux actuels. Les outils MongoDB et Spark nous ont ainsi permis de traiter nos deux jeux de données et de comparer deux grandes plateformes de VAD.

Source

Statista Infographies. « Info graphie: La course aux abonnés des géants de la VOD ». Consulté le 16 novembre 2021, à l'adresse https://fr.statista.com/infographie/24389/evolution-nombre-abonnes-payants-svod-plateformes-streaming-video-netflix-prime-video-disney-/.

Wilmer, J. (s. d.). WATER. FilmFreeway. Consulté le 16 novembre 2021, à l'adresse https://filmfreeway.com/900253

World WarII Foundation. (s. d.). Tim Gray. /wwiifoundation.org. Consulté le 16 novembre 2021, à l'adresse https://wwiifoundation.org/bio-tim-gray/