ALLIX Nicolas
ARSIC Marko
LUTZ Rindra
ROBACHE William

TWITTER : L'IA ET SES LEADERS D'OPINION

Résumé

L'exercice auquel nous nous sommes livrés consistait à collecter des données sur les utilisateurs twitter, afin d'identifier ceux qui représentent des leaders d'opinion sur le hashtag IA. Pour mener à bien cet exercice, nous avons projeté d'utiliser l'API Twitter grâce à un script python que nous avons développé. Cela étant, le peu de temps dont nous disposions ne nous a pas permis de récupérer un nombre de données suffisant, c'est pourquoi nous avons donc décidé d'exploiter les données fournies. Grâce à l'outil Gephi, nous avons réussi à distinguer les profils stratégiques sur le #IA. D'autres graphes ont été réalisées via Tableau afin de présenter nos données de façon plus claire.

Introduction

Nous nous sommes concentrés en groupe sur cette étude lors d'un data challenge enrichissant et créatif. Le but de l'étude qui nous a été donnée est de déterminer parmi une population donnée le leader d'opinion le plus influent sur Twitter. Pour ce faire, nous nous sommes concentrés sur les tweets présentant le hashtag "IA", afin d'orienter l'étude sur le domaine de l'intelligence artificielle. L'enjeu est de profiler le plus possible les leaders d'opinion identifiés, étant donné que les entreprises ont besoin de personnes influentes sur les réseaux sociaux afin de diffuser leurs messages à la plus grande communauté possible.

Critères

Afin de réaliser cette étude sans accroc, nous avons pensé à l'avance aux critères dont nous avions besoin concernant l'identification des leaders d'opinion. Ces critères sont les suivants :

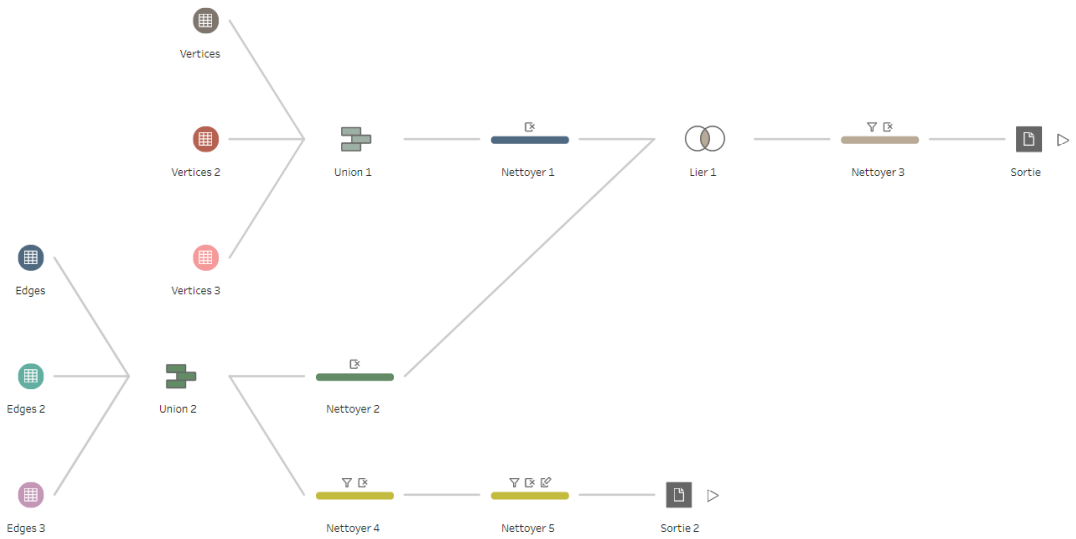
- Nombre de tweets
- Nombre de followers
- Nombre de retweets
- Nombre de publications dans le temps
- Date d'inscriptions

Ces critères nous ont semblé les plus performants et les plus explicatifs concernant l'identification des leaders d'opinion.

Présentation du jeu de données

Après les difficultés techniques auxquelles nous avons fait face, nous avons utilisé les données fournies pour le challenge.

Nous avons utilisé l'ETL Tableau prep pour nettoyer les jeux de données. Dans un premier temps nous avons fusionné les différentes bases de données pour répondre à notre problématique. Nous avons créé deux bases de données en csv. Une première Tweet_analyse.csv ou nous avons gardé le nom des utilisateurs, la date des tweets, le nombre de followers et le nombre de follow, la date d'inscription afin de pouvoir répondre à notre problématique. Cette base de données nous a permis de faire une analyse détaillée des Tweeto sur Tableau Desktop afin de comprendre l'usage de twitter, et une fois relié avec notre graph comprendre ce qui a permis de rendre le Tweeto plus influent que les autres. Le Graph fait sur Gephi en plus de l'analyse Tableau nous a permis de comprendre comment un Tweeto est devenu influent. La deuxième base de données Gephi.csv nous a permis de faire le graphe.



Méthodes et étapes

Pour effectuer notre analyse, premièrement, nous nous sommes tournés vers l'api twitter, afin d'avoir le choix sur les données que nous souhaitons analyser et d'avoir les données en temps réel.
Pour cela, nous avons programmé nos requêtes sur Jupyter, grâce au langage de programmation Python et les librairies : Tweepy (permet d'avoir accès à l'api twitter), Pandas et Numpy pour la mise en forme des données.
Malheureusement, en raison d'une connexion internet trop faible et des ordinateurs pas assez puissants, nous avons atteint la limite d'export d'une semaine, trop faible pour ce que nous souhaitions faire.

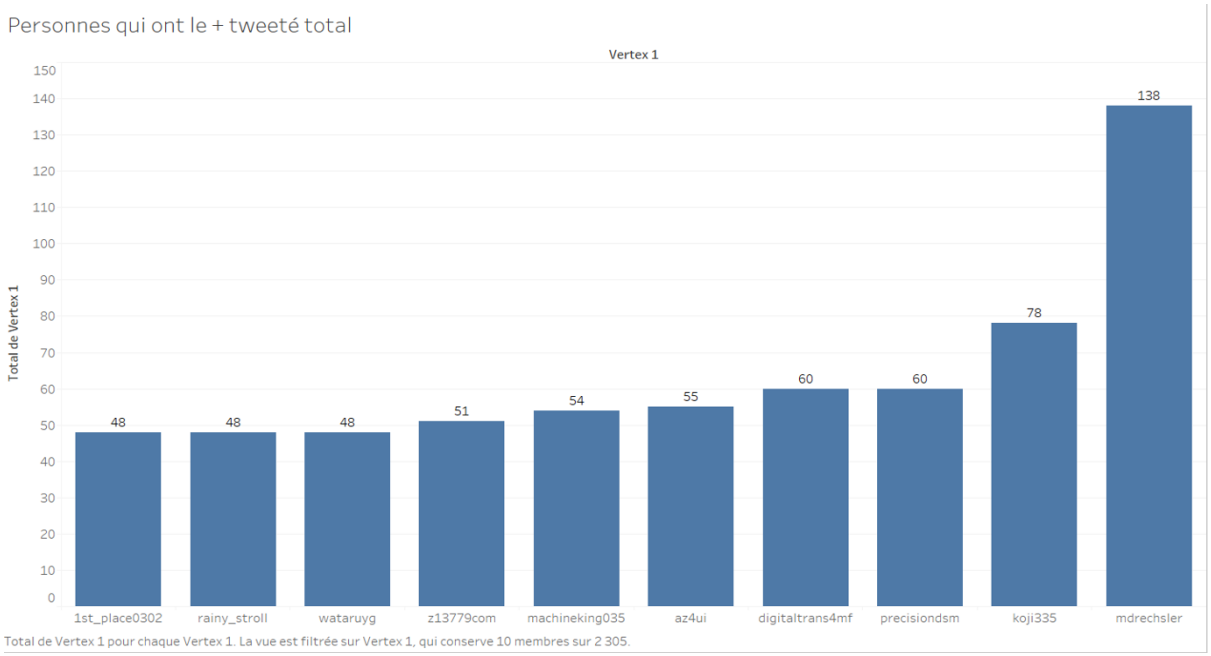
Face à cette limite, nous avons décidé d'utiliser les BDD fournies, pour obtenir une analyse plus précise.

Afin de réaliser notre étude, nous nous sommes séparés en 2 groupes distincts pour optimiser notre temps.
Un groupe s'est occupé de l'analyse pure des BDD à l'aide du logiciel Tableau, après avoir compilé les 3 jeux de données.
L'autre groupe quant-à-lui, s'est chargé de réaliser le graphique en réseau sur Gephi, afin de mettre en avant les liens entre les utilisateurs grâce aux mentions.
Gephi, nous a permis également d'obtenir les données statistiques liées au graphe que nous avons ensuite analysé sur Tableau.
Pour finir, nous nous sommes réunis pour écrire ce rapport sur google docs.

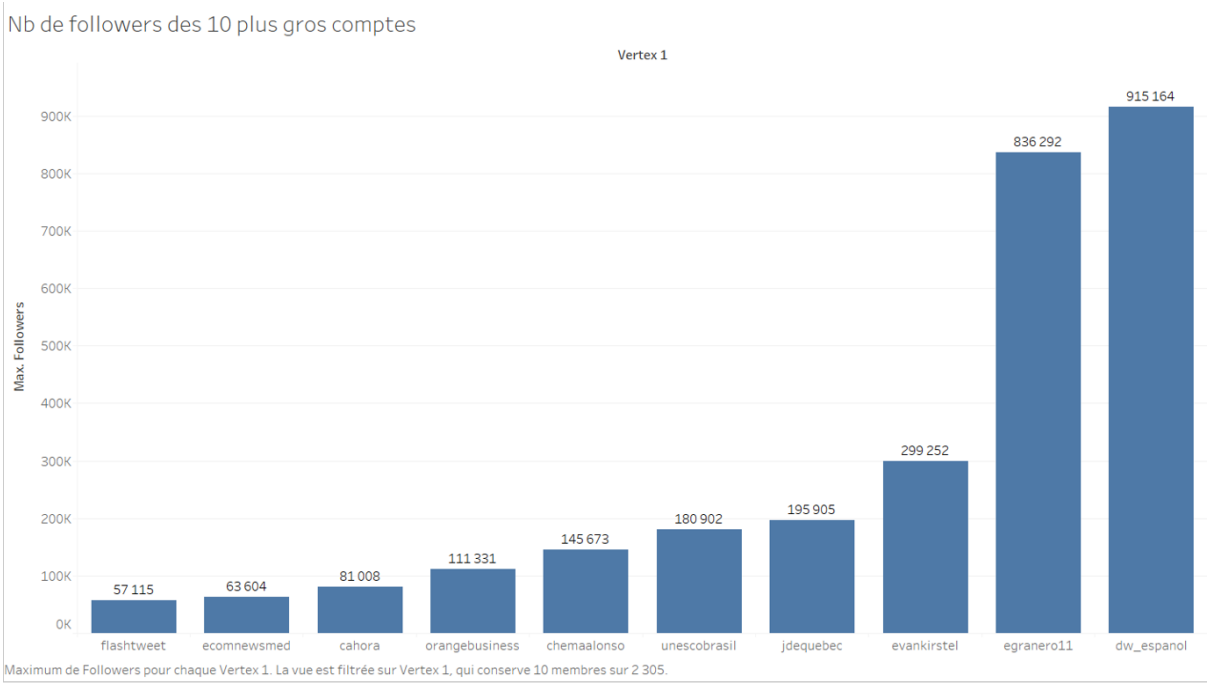
Résultats

Notre analyse nous a permis de mettre en exergue certaines tendances qui ont contribué à reconnaître les leaders d'opinions sur les réseaux sociaux, en particulier sur la twittosphère. Contrairement à ce que nous pourrions penser, les leaders d'opinions ne sont pas les utilisateurs ayant le plus de followers, ou bien ceux qui sont sur la plateforme depuis le plus longtemps. En effet, un des indicateurs les plus parlant est le niveau de contribution soit le nombre de publications qui un élément caractéristique des individus disposant de plus gros pouvoir d'influence sur ce hashtag. Également, le nombre de retweet constitue un élément qui reflète de manière significative la popularité des utilisateurs. Les twittos qui sont le plus actifs, autrement dit dont la fréquence de retweet est la plus élevée ont tendance à disposer d'un plus grand pouvoir d'influence. Ces utilisateurs-là se révèlent comme étant ceux qui disposent de la plus grande capacité de diffusion et de médiatisation.

Dans notre analyse, nous avons confronté les utilisateurs avec le plus de followers aux utilisateurs ayant publié le plus sur une période de 3 semaines.

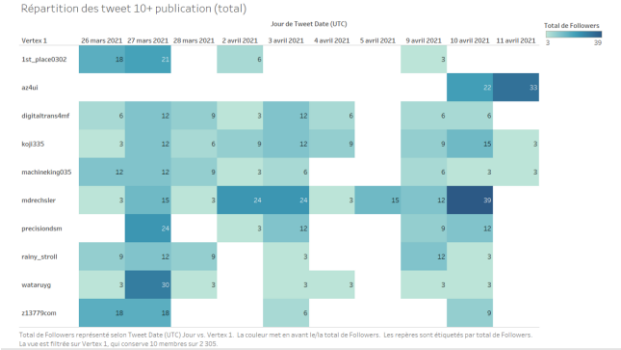


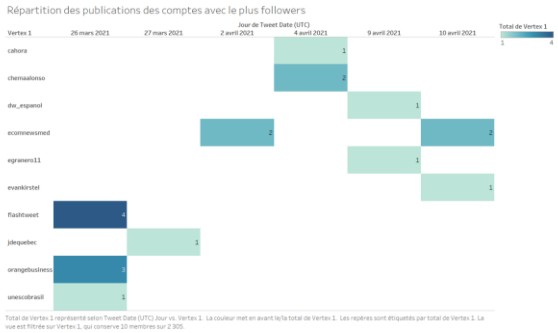
On constate que la personne ayant de loin le plus publié de Tweet est mdrechsier par contre nous avons constaté qu'il n'y a pas de corrélation entre le nombre de tweet publié et le nombre de followers. Nous retrouvons dans le tableau ci-dessous les 10 utilisateurs avec le plus de followers :



Comme expliqué précédemment, aucun des utilisateurs qui publient le plus ne font partie du top 10 des comptes avec le plus de followers.

Au cours des trois dernières semaines, les utilisateurs avec le plus de followers n'ont pas beaucoup publié de post.





La date de création compte n'explique pas la différence de followers avec les utilisateurs avec des publications soutenues.

Date de création des 10 comptes qui publient le plus

Vertex 1	Année de..	Mois de J..	
mdrechsler	2008	novembre	138
koji335	2013	août	78
precisiondsm	2009	mars	60
digitaltrans4mf	2019	février	60
az4ui	2021	avril	55
machineking035	2010	mars	54
z13779com	2013	août	51
wataruyg	2010	avril	48
rainy_stroll	2013	novembre	48
1st_place0302	2011	octobre	48

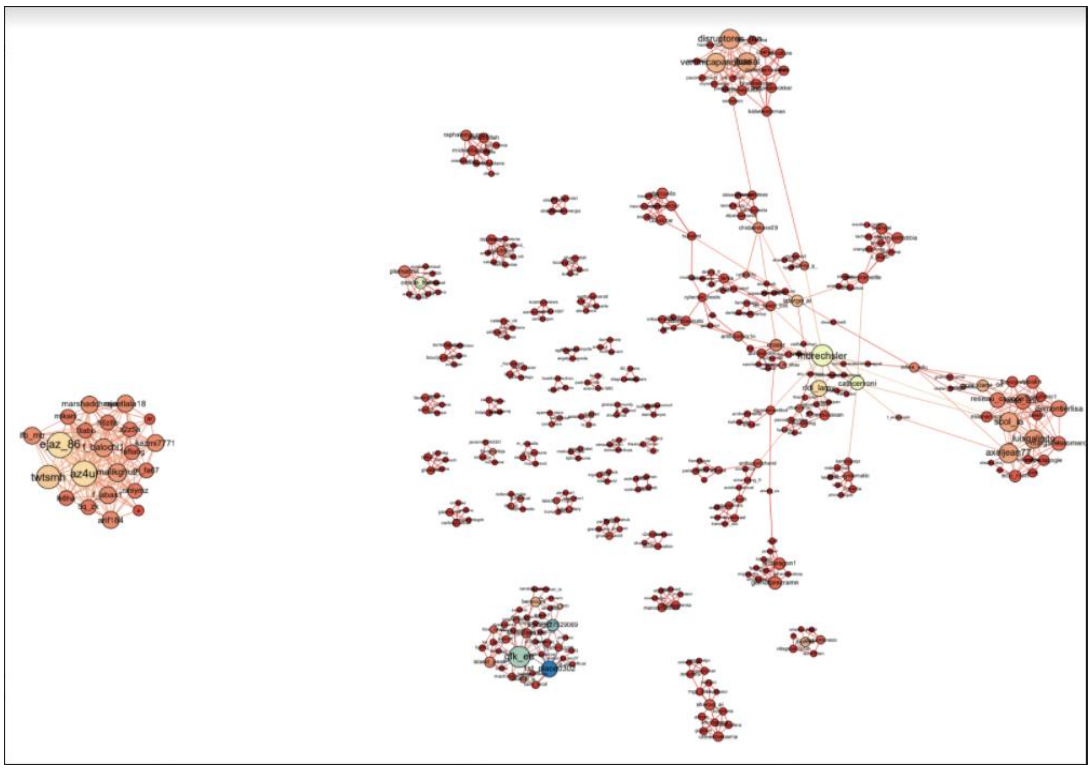
Total de Vertex 1 représenté selon Vertex 1, Joined Twitter Date (UTC) Année et Joined Twitter Date (UTC) Mois. La vue est filtrée sur Vertex 1, qui conserve 10 membres sur 2 305.

Date de création des 10 comptes avec le plus followers

Vertex 1	Année de..	Mois de Joi..	
dw_espanol	2007	mai	915 164
egranero11	2011	février	836 292
evankirstel	2009	avril	299 252
jdequebec	2009	avril	195 905
unescobrasil	2009	août	180 902
chemaalonso	2010	janvier	145 673
orangebusiness	2009	janvier	111 331
cahora	2009	septembre	81 008
ecomnewsmed	2015	septembre	63 604
flashtweet	2010	février	57 115

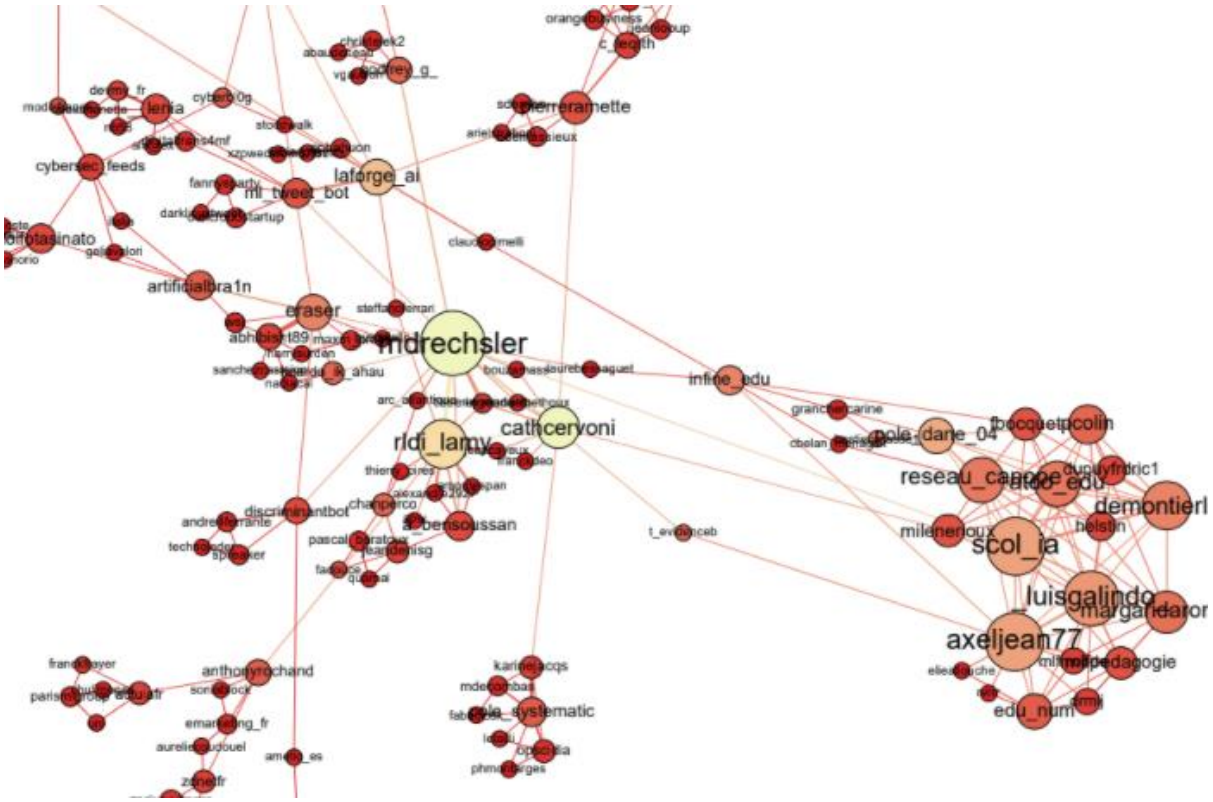
Maximum de Followers représenté selon Vertex 1, Joined Twitter Date (UTC) Année et Joined Twitter Date (UTC) Mois. La vue est filtrée sur Vertex 1, qui conserve 10 membres sur 2 305.

L’analyse des graphes nous montre vers qui le réseau se concentre.

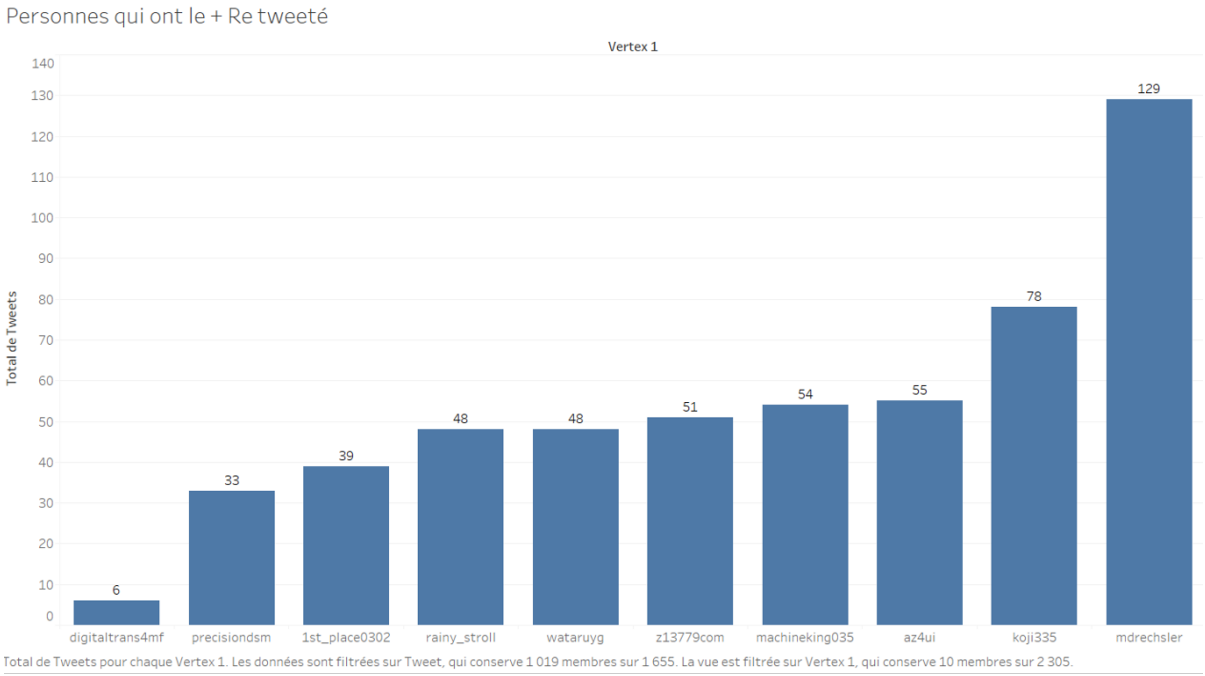


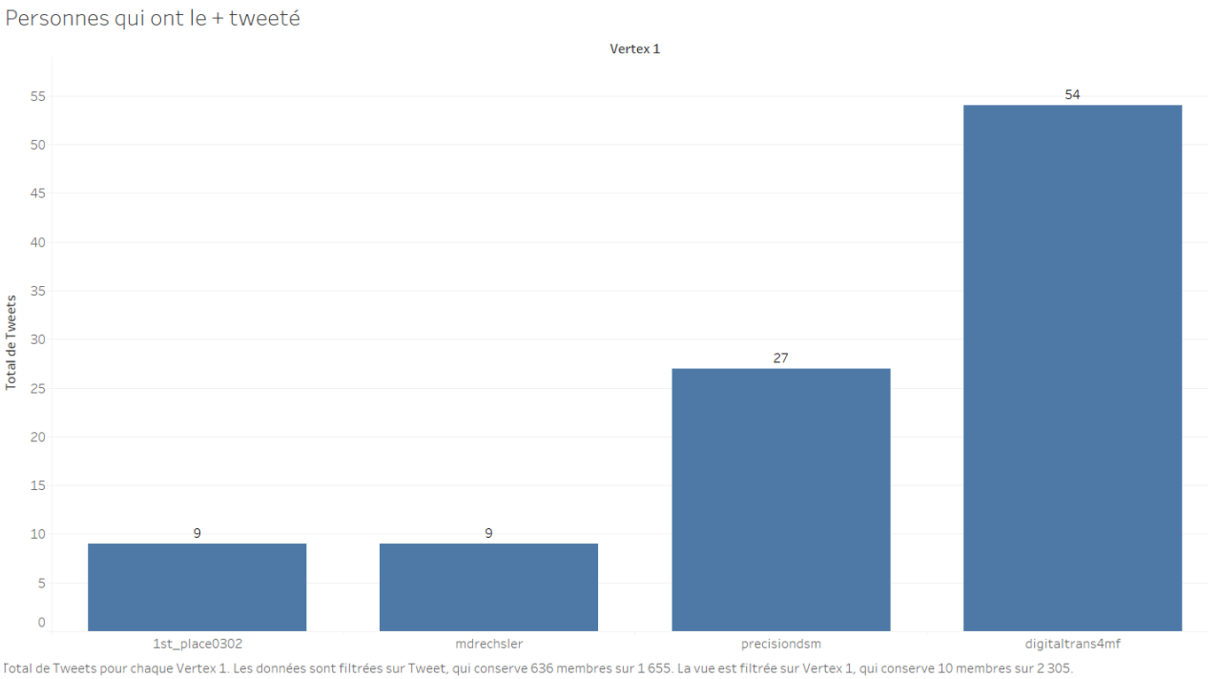
Le graphe nous montre qu'il y a beaucoup de clusters, dont deux assez grands. Enfin, ce graphique nous montre aussi qu'il y a un réseau plus large et plus interconnecté.
La faiblesse des clusters est la diversité des informations partagées mais aussi une audience possible à atteindre qui va mettre plus de temps à se développer.

Au centre du network principal, nous retrouvons encore une fois mdrechsier. Nous en concluons donc que cet utilisateur est le plus influent sur le #IA.



Notre analyse des tweets nous permet de comprendre d'où vient son influence :





Ces deux graphiques nous permettent de comparer le type de post entre ceux créés par l'utilisateur et les retweets.

Nous constatons que peu importe le type de post, le plus important pour gagner en notoriété est de publier le plus souvent possible.

Conclusion

Cette étude a été challengeante à mener. En effet, il nous a fallu réduire en très peu de temps une grande dose de travail en quelques jours, ce qui n'a pas été de tout repos. Nous avons rencontré de nombreuses difficultés, que ce soit en termes d'organisation dans l'urgence, ou bien les limites techniques qui nous ont été imposées. Les bases de données que nous avons essayé de récupérer étaient trop complexes à traiter en termes de poids, et nous n'avons donc malheureusement pas pu les utiliser (vous trouverez en Annexe le travail effectué qui n'a pas pu aboutir). Cet échec nous a poussé à nous rabattre sur les trois bases de données fournies pour le challenge, et nous avons su les exploiter. Grâce à notre travail, il nous a été donné d'observer une tendance concernant le profil type des leaders d'opinion. Ces leaders sont généralement les personnes qui publient le plus sur leur compte, et ont le plus grand nombre de retweet. Les autres critères que nous avons analysés n'ont pas l'air d'influencer grandement le lead d'opinion, ils ne sont donc pas ressortis comme essentiels dans notre analyse. Ainsi, pour parler au plus grand nombre et influencer via les réseaux sociaux, il faudrait que les entreprises diverses communiquent le plus possible à ce type de profil. C'est donc également à ce type de profil qu'il faudra au maximum s'identifier et copier si une tierce personne souhaite devenir un leader d'opinion sur la twittosphère.