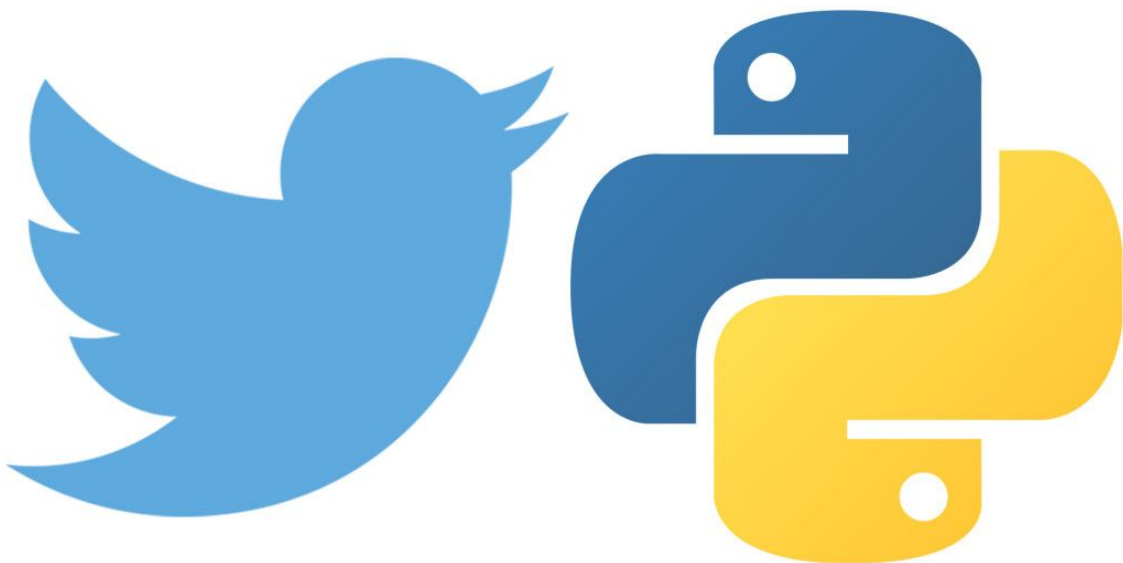


# Logiciel de récupération de données Twitter

Emile Houplain - Nicolas Gendron

Master 2 Stratégies digitales et innovation numérique  
Théorie des réseaux et données relationnelles

---



<b>Introduction</b>	<b>2</b>
<b>I – Description des modules et explication des paramètres</b>	<b>4</b>
A) Module “récupération de données twitter”	4
B) Module “Fusion de bases de données”	7
C) Module “Filtres de bases de donnée”	8
D) Représentation de réseaux	12
E) Module récupération de tweets	14
F) Module filtres de tweets	15
G) Module représentation de tweets	17
<b>II – Exemple d’analyse en utilisant uniquement le logiciel</b>	<b>19</b>
A) Récupération des données du compte @MasterSDIN	19
B) Fusionner des bases de données Twitter	21
C) Filter une base de donnée Twitter	21
D) Représentation de bases de données Twitter.	23
E) Récupération de tweets	25
F) Filtres de tweets	26
G) Représentation de tweets	27
<b>Conclusion</b>	<b>29</b>

# Introduction

Aujourd'hui, les réseaux sociaux, par leur utilisation abondante et variée, que ce soit pour communiquer, réseauter, rechercher de nouveaux partenaires, se créer une vitrine, sont devenus un atout considérable dans la stratégie digitale notamment pour les entreprises. C'est donc un enjeu important que de pouvoir analyser ces réseaux sociaux, or les dynamiques sur ces réseaux sont complexes car de nombreux agents sont en relation de manière continue, ce qui rend l'analyse de ces réseaux difficile.

Cependant il est possible d'effectuer des analyses de ces réseaux, facilité par les APIs (application programming interface) mises en place par les réseaux sociaux eux-mêmes et qui permettent donc la récupération de données sur leurs applications de manière contrôlée. Or pour ce faire il faut maîtriser un langage de programmation, comme python, et cela représente un travail conséquent et donc beaucoup de temps pour un individu qui n'a pas les connaissances. Nous avons donc à partir de ce constat décidé de créer un outil intuitif permettant à un utilisateur d'effectuer des récupérations de données sur Twitter, ainsi que la construction de réseaux relationnelles, et d'autres fonctionnalités additionnelles permettant une analyse plus poussée, notamment une analyse des tweets en fonction de certains critères. L'objectif ici est de permettre à un individu qui n'a pas nécessairement des connaissances approfondies en langages informatiques de pouvoir à travers un logiciel et une interface simple d'utiliser certaines fonctionnalités permises par l'API de Twitter et donc de pouvoir récupérer et analyser un réseau twitter sans avoir à écrire des lignes et des lignes de codes. Nous fournissons donc un outil permettant la construction et l'analyse de réseaux Twitter pour tous.

Ce projet est donc sous la forme d'un petit logiciel programmé grâce à Python, langage de programmation notamment adapté à la récupération de donnée sur internet, c'est d'ailleurs le langage utilisé pour communiquer avec l'API Twitter (appelé Tweepy). Cette dernière facilite la communication avec les serveurs de Twitter, elle permet de récupérer relativement simplement des données issues de comptes Twitter. Cependant, il faut un compte développeur Twitter afin de posséder les identifiants (tokens) nécessaires à l'identification permettant d'utiliser cette API, ces identifiants permettent notamment à Twitter de suivre notre activité de récupération de données. Par exemple, l'API bloquera notre algorithme de

récupération de données twitter au bout d'un certain moment, pour reprendre par la suite, le but étant de ne pas surcharger les serveurs de twitter.

Pour la représentation des réseaux twitter, on utilise la librairie NetworkX, qui permet de dessiner et de représenter les nœuds et les liens, et de les placer en fonction de certains critères. Il est essentiel à l'analyse de réseaux, puisque c'est à partir de ce dernier que sera représenté les réseaux finaux de manière visuelle, et qui permettra donc une meilleure lisibilité de ces derniers. La librairie NetworkX permet également les calculs de différentes mesures d'importances. Enfin, pour créer la visualisation de ce logiciel, c'est à dire pour créer l'intégralité des fenêtres, boutons, c'est à dire de l'interface utilisateur, nous avons utilisé la librairie Tkinter. C'est grâce cette librairie que l'utilisateur va pouvoir naviguer entre les différentes fonctionnalités, spécifier les paramètres qu'il souhaite, et ainsi personnaliser son analyse de réseau en fonction de ce qu'il veut montrer.

Ce logiciel est construit de manière modulable, ce qui veut dire que le logiciel n'est pas figé et donc il est possible d'ajouter des modules supplémentaires. Il est assez simple pour n'importe qui d'ajouter une fonctionnalité (c'est à dire un module) ou un paramètre. De plus les modules sont liés entre eux : l'output de l'un peut être l'input de l'autre, par exemple les bases de données récupérées seront ensuite fusionnées dans le module correspondant. Cela est beaucoup plus adapté à l'utilisateur qui peut directement décider du module à utiliser à partir d'une seule et même fenêtre, en cliquant simplement.

Cet outil est donc divisé en différents modules, accessibles depuis la première fenêtre au lancement du logiciel. Chaque module permet certaines fonctionnalités :

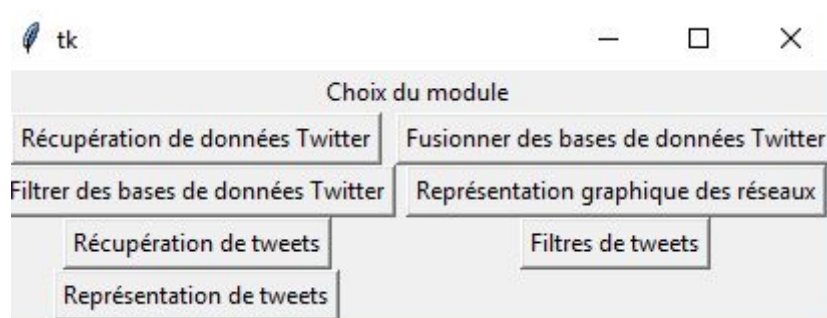
- Le premier permet la récupération de données Twitter (données sur les followers du compte indiqué),
- Le second permet de fusionner différentes bases de données Twitter récoltées auparavant,
- Le troisième permet de filtrer ces bases de données en fonction de différentes conditions
- Le quatrième permet d'effectuer une représentation visuelle du réseau constitué à partir de bases de données Twitter,
- Le cinquième module permet de récupérer les tweets d'un compte twitter spécifique
- Le sixième module permet de filtrer ces tweets récupérés, en fonctions de différents critères,

- Le dernier module permet de représenter une base de données de tweets en tableur et en carte géographique.

Dans une première partie nous décrirons les différents modules et nous expliquerons les différents paramètres disponibles sur ce logiciel, puis dans une seconde partie nous mettrons en application notre logiciel afin de comprendre son fonctionnement de manière concrète.

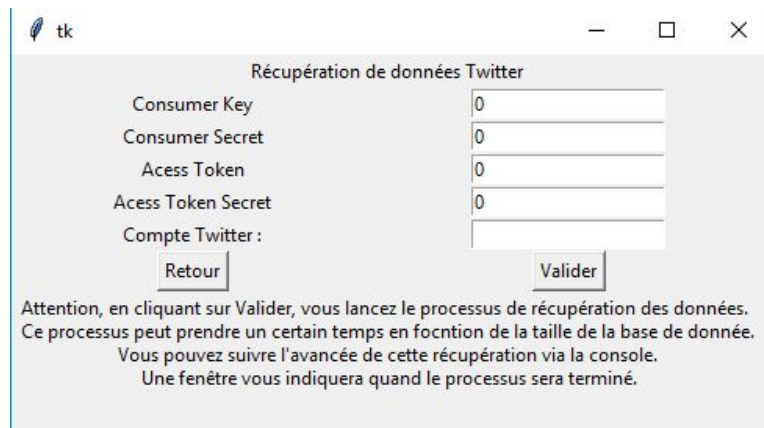
## I – Description des modules et explication des paramètres

Tout d’abord, au lancement du logiciel (fichier Tkinter\_gen), une fenêtre d'accueil s’ouvre, qui sert d’index pour naviguer entre les différents modules :



### A) Module “récupération de données twitter”

Si l'utilisateur clique sur le bouton “Récupération de données Twitter” une nouvelle fenêtre s’ouvrira :



tk

Récupération de données Twitter

Consumer Key 0

Consumer Secret 0

Access Token 0

Access Token Secret 0

Compte Twitter :

Retour Valider

Attention, en cliquant sur Valider, vous lancez le processus de récupération des données.  
Ce processus peut prendre un certain temps en fonction de la taille de la base de données.  
Vous pouvez suivre l'avancée de cette récupération via la console.  
Une fenêtre vous indiquera quand le processus sera terminé.

Le premier module est indispensable à l'analyse de réseaux twitter puisqu'il permet de récupérer les données qui seront nécessaires à la construction des réseaux et à leur analyse. Pour cela, le logiciel utilise l'API Tweepy mise à disposition par Twitter afin de récupérer les données des comptes twitter voulus.

Il faut donc un compte développeur twitter (<https://developer.twitter.com>) afin d'obtenir les identifiants nécessaires pour être autorisé à récupérer des données sur twitter via l'API. Ces identifiants sont au nombre de quatre et sont disponibles sur votre compte développeur.

Ces identifiants devront être entrés dans les champs correspondant du module de récupération de données, le nom du compte tweeter à entrer est le nom original du compte twitter, c'est à dire le @ (exemple de compte à entrer dans le champ : @exemple). Il suffira ensuite de cliquer sur "Valider" pour lancer la récupération de données twitter du compte voulu.

Les données récupérées sont en fait les données de tous les followers du compte analysé : leurs nombres de tweets, leurs nombres d'amis (friends), d'abonnements (followers), mais également leurs listes de followers respectives, le but étant potentiellement d'établir les liens entre ces comptes twitter. C'est donc une analyse du réseau des followers du compte étudié.

Cependant, cette récupération peut être longue, en fonction du nombre de followers du compte, et de leurs followers respectifs (puisque le logiciel récupère les followers des followers du compte twitter). La récupération des données d'un compte twitter ayant des followers « importants », c'est à dire eux-mêmes très suivis. Cette récupération peut prendre quelques minutes pour les petits comptes, à plusieurs jours pour les bases les plus

conséquentes. Il est possible de suivre l'avancée de cette récupération de donnée via la console s'ouvrant au lancement du logiciel. Face à ces problèmes de temps de récupération de données, il est possible d'effectuer cette dernière en plusieurs fois, ainsi si les données récupérées sont enregistrées sur l'ordinateur de l'utilisateur au fur et à mesure. Au lancement d'une récupération de donnée, le logiciel va automatiquement chercher s'il y a déjà une base de données à ce nom existante, puis il va continuer la récupération de donnée si c'est le cas. C'est pour cela que le fichier de sortie de la base de données récupérée est automatiquement enregistré à l'adresse @CompteTwitter/Datas\_globales\_@CompteTwitter et qu'il est fortement recommandé de ne pas le changer de place. D'une part pour pouvoir reprendre la récupération de donnée si elle n'a pas été finie auparavant, mais également pour garder la base de données brute non modifiée à l'aide de filtre ou autre, pour pouvoir revenir en arrière si besoin. De manière générale il est recommandé de garder trace de chaque étape de modification de la base de données.

Cette récupération de données peut être d'autant plus longue que l'API Twitter suit l'activité de récupération de donnée et bloque cette dernière en fonction d'un "quota" pour éviter de surcharger les serveurs en requêtes de récupération de données. La vitesse est donc bridée par l'API Twitter elle-même.

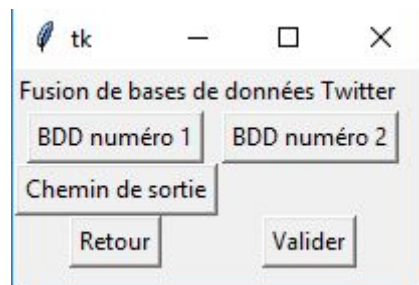
De plus, il est possible que certaines données devant être récupérées soient "bloquées". En effet il est possible pour un compte tweeter de bloquer l'accès à une partie ou à l'intégralité de ses données, dans ce cas, la récupération de données relatives à ce compte sera biaisée. Si c'est le cas, le compte twitter correspondant dans la base de données sera indiqué comme ayant un problème et sera supprimé lors d'application de filtres ou de la création de réseaux.

A partir de la base de données récoltée, des liens sont construits entre les comptes de cette base de données, c'est à dire les comptes en relation sur twitter. Si entre deux comptes twitter A et B l'un au moins suit l'autre, il y a une relation entre les deux, et un lien est créé entre ces deux comptes. L'ensemble des liens sont présents dans la base de données enregistrée. Il faut cependant prendre en compte une des limites de cette méthode de construction de liens : ces liens ne sont pas dirigés. En effet, il n'est pas pris en compte le sens de la connexion entre deux comptes (on ne prend pas en compte le fait que ce soit le compte A qui suit le compte B, ou le compte B qui suit le compte A, on prend seulement en

compte le fait qu'il y ait une liaison entre ces deux comptes).

## B) Module "Fusion de bases de données"

Si l'utilisateur clique sur le bouton "Fusionner des bases de données Twitter" à partir de l'index, la fenêtre suivante s'ouvre :



Une fois le premier module utilisé pour récupérer les données des comptes twitter voulus, il est possible de fusionner ces bases de données, pour pouvoir ensuite analyser cette nouvelle base fusionnée dans sa globalité. Ce module permet cela en indiquant les deux chemins d'entrée des deux bases à fusionner, et le chemin de sortie de la future base fusionnée.

Il est possible de fusionner des bases déjà fusionnées, ou filtrées auparavant, comme des bases de données globales. Cela permet notamment une analyse plus générale d'un domaine, par exemple, pour analyser le domaine de la RSE en Bretagne, il est intéressant de récupérer les données relatives à plusieurs comptes twitter sur le thème de la RSE localisés en Bretagne, de les fusionner en une seule et même base de données pour ensuite l'analyser, plutôt que de récupérer uniquement les données relatives au réseau d'un seul compte twitter. Cela permet donc une analyse plus large et plus donc souvent plus pertinente.

Cependant comme expliqué précédemment, plus les bases de données seront nombreuses et conséquentes à récolter plus cela prendra du temps, il faut donc effectuer un arbitrage entre le coût en termes de temps et le gain en termes de qualité de représentation des réseaux.



Suite à la fusion de deux bases de données, les doublons présents seront supprimés. En effet il est possible qu'un même compte twitter soit présent dans deux bases de données différentes s'il est en relation avec les comptes twitter dont on a récupéré les données (liste des followers, données de ses followers ...).

Il faut cependant faire attention à un phénomène : si deux bases de données d'un compte A et B sont fusionnées (donc les données relatives aux followers des comptes A et B), et que A est présent dans les followers de B (ou inversement), alors A sera présent dans la base de données fusionnée, et sera relié par conséquent à ses followers, également présents dans la base de données.

Enfin, les liens seront mis à jour pour correspondre à la nouvelle base de données fusionnée afin de savoir quel compte est en relation avec quel compte, en vue de pouvoir filtrer cette base de données, ou directement construire le réseau de cette dernière.

Cette nouvelle base de données fusionnée sera enregistrée dans un dossier spécifié par l'utilisateur. Elle pourra être de nouveau utilisée pour être filtrée, ou une nouvelle fois fusionnée avec une autre base de données. Il est également possible de directement la représenter sous forme de réseau grâce au module correspondant, même s'il est recommandé de filtrer auparavant les bases de données pour une meilleure lisibilité afin de faire ressortir ce qui est intéressant à analyser.

## C) Module "Filtres de bases de donnée"

Si l'utilisateur clique sur le bouton "Filtrer des bases de données Twitter", la fenêtre suivante s'ouvrira :

tk

Filtre de bases de données Twitter

Base de donnée à filtrer

Filtrage par nombre de Tweets : ☒

Nombre de Tweets : 0

Filtrage par nombre de Friends : ☒

Nombre de Friends : 0

Filtrage par nombre de Followers : ☒

Nombre de Followers : 0

Filtrage par importance du compte (en terme de mesure) ☒

Choix de la mesure : Degré de centralité  
Closeness  
Eigen Vector

Pourcentage des comptes les moins importants supprimés : 0.00

Ecriture de la base en fichier CSV : ☐

Chemin de sortie

Retour Valider

Ce module permet d'appliquer des filtres en fonction de certains critères. Les comptes twitter de la base de données filtrée ne correspondant pas aux critères seront supprimés de la nouvelle base de données. Cela permet de supprimer les comptes qui ne paraissent pas pertinent aux yeux de l'utilisateur en fonction de ce qu'il veut analyser. Par exemple, si une entreprise cherche à identifier des comptes twitter qui semblent influents dans un certain domaine, il est intéressant de supprimer les comptes qui ne semblent pas importants en termes de followers. De même, les comptes qui semblent intéressants pour communiquer sont les comptes actifs, et donc qui tweets relativement souvent. Grâce à ces filtres il est donc possible de supprimer de la base de données les comptes non pertinents.

Le premier filtre qui sera appliqué automatiquement à la base de données est la suppression des comptes ayant eu un problème lors de la récupération de données. Cela est obligatoire car pour appliquer les autres filtres, il faut avoir accès à ces données.

Le deuxième filtre pouvant être appliqué est le nombre de tweet. Si l'utilisateur indique qu'il faut appliquer un filtre par rapport au nombre de tweets, l'ensemble des comptes twitter de la base de données ayant moins que le nombre de tweets spécifié par l'utilisateur seront supprimés de la base de données. Il faut bien prendre en compte le fait qu'un retweet (c'est à dire le faire de partager un tweet de quelqu'un d'autre sur son profil) est compté comme un tweet.

Le troisième filtre concerne le nombre d'amis (friends) du compte twitter, c'est à dire le nombre de comptes twitter suivi par ce compte. Si ce filtre est activé, les comptes de la base de données ayant moins d'amis que le nombre spécifié par l'utilisateur seront supprimés. Ce filtre est relativement peu pertinent dans les études d'importance puisqu'il traduit l'idée de comptes suivis, et non de followers. En effet un compte twitter peut très bien s'abonner à de nombreux autres comptes twitter importants en termes d'influence, mais n'être lui-même que très peu suivi, ainsi ses tweets auront sûrement peu d'influence. A l'inverse, un compte ayant beaucoup de followers importants peut avoir peu d'amis, c'est à dire de comptes suivis. C'est le cas par exemple du compte twitter de l'actuel président américain Donald J. Trump (@realDonaldTrump) qui possède 45 amis, pour 57 millions d'abonnés, c'est à dire de followers. Il faut donc faire attention avec l'utilisation de ce filtre qui peut rapidement supprimer des comptes qui peuvent être pertinents pour une analyse d'importance.

Le quatrième filtre s'applique de la même façon que le précédent, mais au nombre de followers. Il est intéressant à utiliser pour supprimer des comptes non influents. En effet, un compte ayant peu de followers n'aura que peu d'influence, ses tweets auront moins de chances d'être vus, donc d'être retweeté. Cependant il faut faire attention car ce filtre ne prend pas en compte l'importance des followers eux-mêmes. Ainsi un compte qui pourrait être important car suivi par certains followers importants pourrait être supprimé de la base de données.

Le cinquième filtre concerne les mesures d'importance. En effet il est possible de calculer pour chaque nœud, c'est à dire pour chaque compte twitter, une "note d'importance ou d'influence" en fonction de ses relations au sein du réseau, c'est à dire de ses liens aux autres comptes de la base de données. Il est donc possible de supprimer x% des comptes de la base de données les moins importants selon une mesure précise. Les mesures qu'il est possible d'utiliser sont au nombre de trois et représentent chacune un aspect différent de la notion d'importance dans un réseau.

Pour rappel, un nœud est un compte twitter, un lien entre deux nœuds représente une relation, c'est à dire le fait qu'au moins un des deux comptes suivent l'autre.

La première mesure est la plus fréquente et la plus intuitive : c'est le degré de centralité. Le degré de centralité correspond au nombre de liens incidents à un nœud (c'est-à-dire le

nombre de voisins que possède un nœud). Il représente concrètement le pourcentage de nœuds du réseau auxquels il est relié, il ne prend donc pas en compte l'importance des nœuds. En effet, être relié à un nœud "important" (c'est à dire lui-même relié à beaucoup d'autres nœuds), ou être relié à un nœud peu important (c'est à dire peu ou pas relié à d'autres nœuds) ajoute la même valeur au degré de centralité.

La deuxième mesure utilisée pour analyser le réseau constitué est la mesure "closeness" qui traduit l'idée d'éloignement moyen par rapport aux autres nœuds du réseau. En moyenne, pour atteindre un nœud du réseau, par combien d'autres nœuds doit-on passer, en prenant en compte les relations entre les nœuds. Cette mesure est très intéressante pour étudier une dynamique de diffusion de l'information. Elle permet en effet de mesurer la force de connexion d'un nœud par rapport à tous les autres nœuds du réseau (et non pas seulement aux nœuds reliés au nœud analysé). Un compte ayant un degré de centralité très faible (relié uniquement à un seul nœud du réseau), peut avoir une forte valeur de "closeness" car le seul nœud auquel il est relié, est relié à tous les autres nœuds du réseau.

La troisième mesure, la mesure "eigen vector", prend en compte la centralité des nœuds voisins, c'est à dire les comptes ou nœuds reliés au nœud étudié. Cette mesure permet donc de corriger la limite de la première mesure évoquée précédemment.

A noter que le pourcentage représente le pourcentage des comptes les moins importants à supprimer, par exemple si l'utilisateur indique 0.4 en pourcentage, ce seront les 40% des comptes les moins importants (en fonction de la mesure choisie) qui seront supprimés de la base de données, et les 60% les plus importants seront gardés dans la base de données.

Il est également possible d'écrire la nouvelle base de données filtrée dans un fichier CSV, qui permet une certaine lisibilité des données. Le fichier CSV sera enregistré au même endroit que le chemin de la base de données filtrée spécifié par l'utilisateur et portera le même nom.

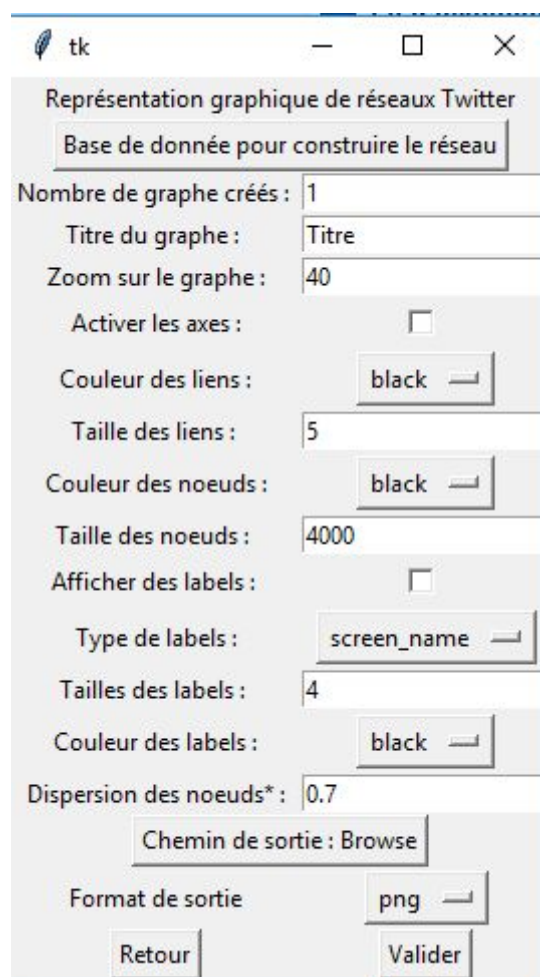
Cependant les filtres proposés restent simples, et sont donc relativement limités. Il serait intéressant de proposer un panel de filtres supplémentaires, avec par exemple la possibilité de supprimer les comptes twitter n'ayant pas tweeté récemment (en fonction d'une date), ce qui permettrait de supprimer les comptes importants selon nombre de tweets, de relations

ou de followers, mais inactifs.

Une fois la base de données filtrée, elle sera donc enregistrée à l'endroit spécifiée par l'utilisateur. Elle pourra être utilisée pour être de nouveau filtrée, fusionnée à une autre base de données, ou encore pour construire le réseau de cette dernière de manière visuelle pour l'analyser.

## D) Représentation de réseaux

Si l'utilisateur clique sur le bouton "Représentation graphique de réseaux" à partir de la fenêtre d'index permettant la navigation entre les différents modules, l'utilisateur verra apparaître une fenêtre lui demandant certains paramètres :



The image shows a Tkinter window titled "Représentation graphique de réseaux Twitter". The window contains several input fields and buttons for configuring a network graph. The parameters are as follows:

- Base de donnée pour construire le réseau**: A button to select the data source.
- Nombre de graphe créés**: A text input field with the value "1".
- Titre du graphe**: A text input field with the value "Titre".
- Zoom sur le graphe**: A text input field with the value "40".
- Activer les axes**: A checkbox that is currently unchecked.
- Couleur des liens**: A color selection button showing "black".
- Taille des liens**: A text input field with the value "5".
- Couleur des noeuds**: A color selection button showing "black".
- Taille des noeuds**: A text input field with the value "4000".
- Afficher des labels**: A checkbox that is currently unchecked.
- Type de labels**: A dropdown menu showing "screen\_name".
- Tailles des labels**: A text input field with the value "4".
- Couleur des labels**: A color selection button showing "black".
- Dispersion des noeuds\***: A text input field with the value "0.7".
- Chemin de sortie**: A button labeled "Browse" to select the output file path.
- Format de sortie**: A dropdown menu showing "png".
- Retour**: A button to go back to the previous screen.
- Valider**: A button to confirm the settings and generate the graph.

L'objectif de ce module est de permettre une représentation graphique du réseau des comptes twitter de la base de données, c'est à dire des relations et des liens qu'ils ont entre eux. Cela facilitera la visualisation de ce réseau, et permettra donc une meilleure analyse et compréhension.

Pour cela, on utilise la librairie NetworkX qui permet de construire ces réseaux en fonction de différents paramètres. Ainsi, l'utilisateur pourra partiellement personnaliser la représentation de son réseau en fonction de ses besoins.

Il est souvent important d'appliquer auparavant des filtres pour réduire le nombre de comptes twitter présents dans la base de données, et ne garder que ceux qui sont intéressants à analyser. Cela permettra de représenter uniquement les comptes intéressants de manière graphique et ne surchargera pas la représentation du réseau. Cela permettra donc une meilleure lisibilité, une meilleure analyse et une meilleure compréhension du réseau.

Tout d'abord l'utilisateur doit renseigner la base de données qu'il souhaite représenter. Ensuite il faut renseigner le nombre de graphes que l'on veut établir, c'est à dire le nombre de représentation visuelle. En effet, il est souvent intéressant d'effectuer plusieurs représentations du même réseau avec différents paramètres : les représentations seront toutes visuellement différentes, mais correspondront toutes au réseau, seulement certaines seront plus lisibles, et plus pertinentes selon ce que l'utilisateur veut montrer.

Le titre du graphe est le titre qui sera affiché sur la représentation visuelle.

Le zoom permet d'ajuster le zoom sur la représentation visuelle du réseau. Un réseau avec de nombreux nœuds et de nombreux liens aura tendance à être plus grand et sera moins lisible.

L'utilisateur peut également choisir d'activer ou non les axes (X et Y) sur le graphique selon ses besoins.

Au niveau des liens, l'utilisateur peut personnaliser leur couleur, leur taille (c'est à dire leur largeur) pour que son réseau soit plus lisible. Des liens trop larges ou des couleurs mal

adaptées peuvent rendre difficile la lecture du graphique.

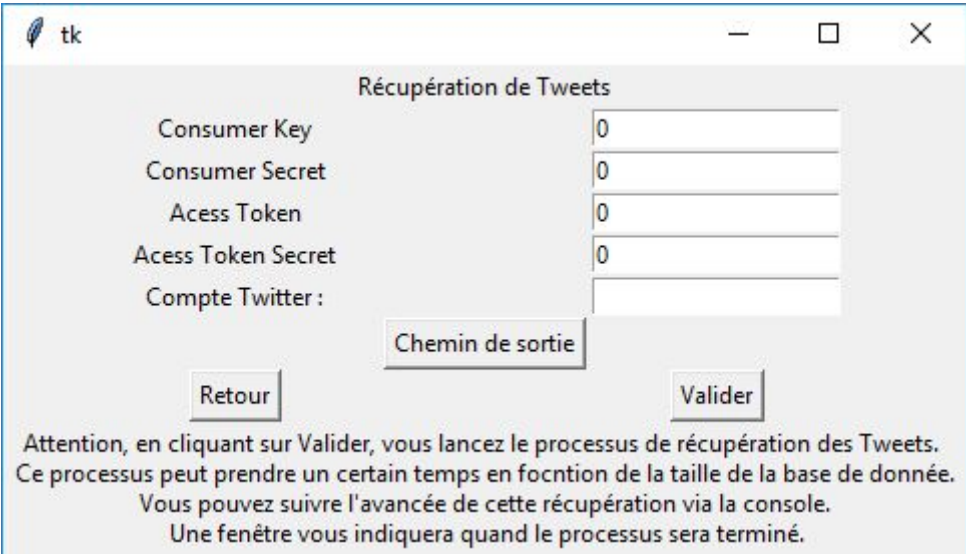
Les labels sont liés aux nœuds et sont représentés au-dessus de ces derniers. Ils peuvent correspondre au nom du compte twitter “original” (c'est à dire le nom du compte tweeter qu'il n'est pas possible de changer, c'est le @CompteTwitter), mais aussi au “pseudonyme” du compte tweeter (c'est le nom affiché sur Twitter, celui qu'il est possible de modifier). Il est également possible d'utiliser les identifiants des comptes en tant que labels.

L'utilisateur peut également choisir la dispersion des nœuds, c'est à dire l'écart entre ces derniers. Plus ce nombre sera élevé, plus les noeuds seront écartés les uns des autres. Ce paramètre peut être important pour une question de lisibilité.

Enfin, l'utilisateur devra indiquer où il souhaite que le ou les fichiers s'enregistrent ainsi que le format dans lequel il veut que cela soit enregistré. Puisque c'est une représentation visuelle, il est proposé le format pdf et png.

## E) Module récupération de tweets

Ce module est accessible en cliquant sur le bouton “Récupération de Tweets” à partir de la première fenêtre de navigation :



The screenshot shows a Tkinter window titled "Récupération de Tweets". It contains several input fields and buttons. The fields are labeled "Consumer Key", "Consumer Secret", "Access Token", "Access Token Secret", and "Compte Twitter :". Each of the first four fields has a text entry box containing the number "0". Below these fields is a label "Chemin de sortie" with an associated text entry box. At the bottom left is a "Retour" button, and at the bottom right is a "Valider" button. Below the buttons, there is a block of text providing instructions: "Attention, en cliquant sur Valider, vous lancez le processus de récupération des Tweets. Ce processus peut prendre un certain temps en fonction de la taille de la base de donnée. Vous pouvez suivre l'avancée de cette récupération via la console. Une fenêtre vous indiquera quand le processus sera terminé."

Ce module permet de récupérer une liste de tweets d'un compte twitter, et de les stocker dans un fichier, afin de les analyser par la suite. Il peut être intéressant d'analyser les tweets de certains comptes (qui peuvent notamment être identifiés grâce aux modules précédents), notamment afin d'analyser leur dynamique de diffusion.

Ce module fonctionne de la même manière que le module permettant de récupérer les données relatives aux followers d'un compte twitter. L'utilisateur doit renseigner ses identifiants fournis à travers son compte développeur twitter, afin de pouvoir lancer le processus de récupération des tweets. L'utilisateur doit également renseigner le compte twitter dont il veut récupérer les tweets (le compte twitter doit être renseigné avec le nom original de ce dernier : @exemple). De la même façon que pour la récupération des données twitter, le logiciel utilise l'API Tweepy, et peut donc être bloqué par cette dernière lors du processus de récupération de données.

L'utilisateur doit également renseigner le chemin de sortie du fichier enregistré contenant les tweets du compte étudié. Ce fichier pourra être utilisé dans d'autres modules permettant de filtrer ces tweets, et/ou de les représenter de différentes manières.

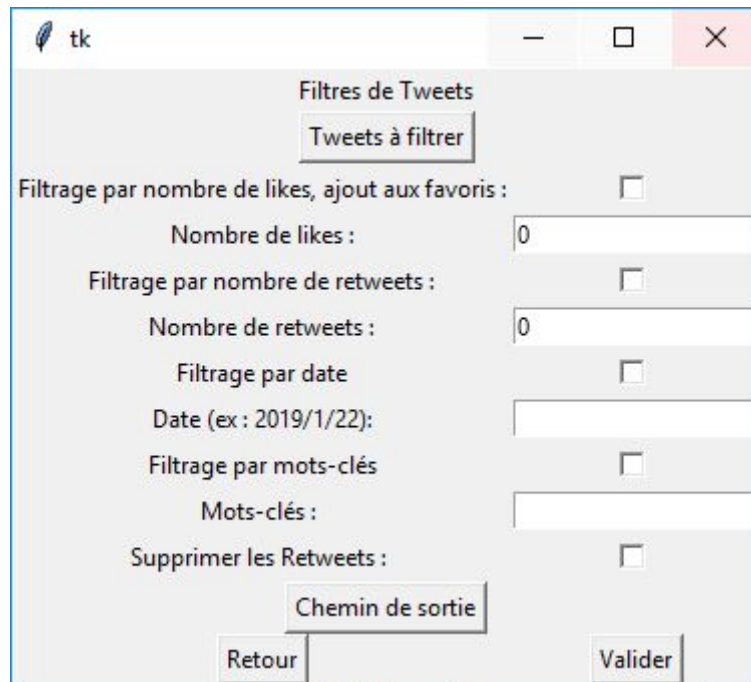
Cependant, ce module ne permet pas de récupérer plus de 4000 tweets d'un même compte twitter, cela peut donc poser des problèmes pour la récupération d'une base de donnée de tweets conséquente.

Il est important de noter que l'utilisateur doit bien différencier une base de données Twitter d'un compte (cette base de données comprend l'ensemble des données relative aux followers d'un compte, c'est à dire leurs followers respectifs, leur nombre de tweets etc.) et les bases de données de Tweets (comprenant la liste des tweets récupérés du compte twitter spécifié : leur contenu, mais également leur nombre de retweet, leur date de publication etc.)

## F) Module filtres de tweets



La fenêtre correspondant au module filtres de tweets est accessible en cliquant sur le bouton "Filtres de tweets" présent sur la fenêtre d'index qui sert à la navigation entre les différents modules.



Une fois le module de récupération de tweet utilisé, il est possible d'utiliser la base de données de tweets afin de l'analyser. Pour cela il peut être intéressant de filtrer ces tweets en fonctions de différents critères pour permettre à l'utilisateur d'effectuer une analyse personnalisée des tweets récupérés.

Ce module fonctionne de la même façon que le module permettant de filtrer les bases de données de compte. L'utilisateur renseigne les tweets qui vont être filtrés, en indiquant le chemin du fichier correspondant à la base de données voulue, puis renseigne les filtres qu'il veut appliquer à ces tweets afin de supprimer ceux qui ne lui paraissent pas pertinents pour son analyse, puis enregistre cette nouvelle base de données filtrée (ne contenant plus que les tweets correspondant aux filtres indiqués) dans un chemin spécifié.

Ces tweets filtrés pourront être utilisés de nouveau, pour être une nouvelle fois filtrés, ou pour être représentés visuellement.

Il est donc possible, grâce à ce module de filtrer les tweets récupérés auparavant en fonction de différents paramètres : nombre de "likes" (c'est à dire nombre de fois que le

tweet a été ajouté aux favoris par un autre compte twitter, qui mesure donc une mesure de pertinence du tweet).

Il est également possible de filtrer ces tweets en fonction de leurs retweets, qui permet une analyse de la diffusion des tweets.

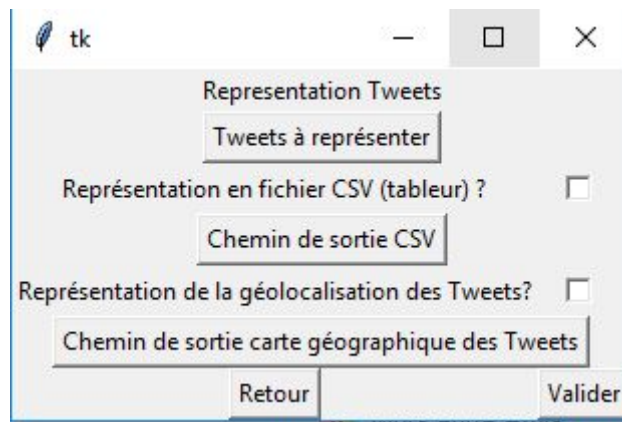
L'utilisateur peut également décider de supprimer certains tweets en fonction d'une date : les tweets publiés avant la date spécifiée seront supprimés de la future base de données de tweets filtrée. Une des limites de ce filtre est qu'il n'est pas possible de choisir un intervalle de tweets afin d'effectuer une analyse comparative (par exemple comparer la fréquence de tweets sur un sujet en fonction de mots clés, entre deux périodes de temps).

Il est également possible de ne garder que les tweets comprenant certains mots clés : l'utilisateur peut renseigner un ou plusieurs mots clés (si plusieurs mots clés : il faut les renseigner séparés par une virgule et sans espace, les tweets gardés seront ceux comprenant au moins un des mots clés renseignés). Ce filtre peut être intéressant pour faire ressortir les tweets abordant certains sujets ciblés.

Le dernier filtre qu'il est possible d'appliquer à une liste de tweets récupérés précédemment est la suppression des retweets. En effet, lors de la récupération des tweets d'un compte, les retweets effectués par ce dernier sont compris dans la liste de ses tweets. Il peut donc être intéressant de supprimer les retweets, pour ne laisser dans la base de données uniquement les tweets publiés originalement par le compte twitter. Cela peut permettre une analyse plus précise, car un compte twitter avec peu d'influence (peu de followers par exemple) ne peut tweeter qu'à travers des retweets (ces retweets ont souvent de bonnes statistiques en termes de nombre de retweet et de like), et donc avoir par exemple une moyenne de retweets par tweet conséquente alors que ce ne sont que des retweets de compte influent.

## G) Module représentation de tweets

De la même manière que pour tous les autres modules, ce module permettant de représenter des bases de données de tweets est accessible depuis la fenêtre de navigation, en cliquant sur le bouton correspondant ("Représentation de Tweets").



Ce dernier module permet la représentation visuelle d'une base de données de tweets (filtrée ou non) récupérée auparavant. Cette représentation peut être sous forme de tableur à travers un fichier CSV, mais également sous forme d'une carte géographique à travers un navigateur internet, en utilisant la librairie Folium.

La représentation de la base de données sous forme d'un tableur permet à l'utilisateur d'avoir une meilleure lisibilité des tweets, et donc une potentielle meilleure analyse à travers les outils disponibles à partir du tableur. Chaque ligne correspond à un tweet, et pour chaque tweet il est renseigné sa date de publication, son nombre de retweets, son nombre de "likes" (ajout aux favoris), son contenu textuel, s'il s'agit d'un retweet ou non, et ses coordonnées géographiques (latitude et longitude) s'il en possède.

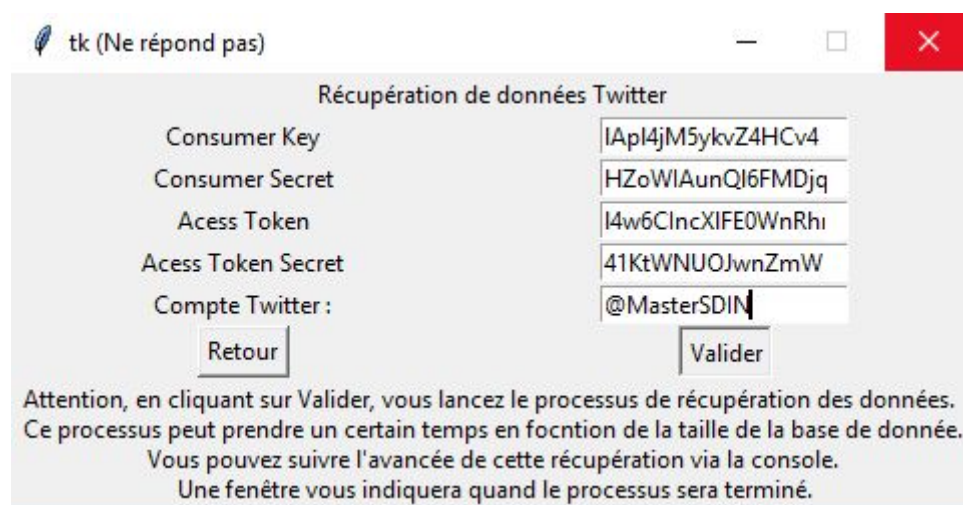
La représentation des tweets sous la forme géographique s'effectue grâce à la librairie Folium et permet une analyse spatiale des tweets. Ces tweets sont représentés sous formes de marqueurs, positionnés à l'emplacement géographique où ils ont été publiés, selon les données récupérées issues de la géolocalisation. Si l'utilisateur clique sur un marqueur, correspondant donc à un tweet, le contenu du tweet s'affiche, ainsi que sa date de publication. À noter que la géolocalisation des tweets est optionnelle et que relativement peu de comptes l'utilisent (par exemple, sur les 12 derniers mois, le compte officiel de Trump n'a jamais publié de tweets géolocalisés). Dans le même ordre d'idée, certains comptes ne l'utilisent que partiellement, et donc certains tweets ne seront pas représentés sur la carte. L'utilisateur peut vérifier la proportion de tweets géolocalisés à partir du tableur (fichier csv : les tweets géolocalisés ont leurs coordonnées renseignées dans la colonne 'géolocalisation').

Il pourrait être intéressant d'ajouter une fonctionnalité permettant de graduer les couleurs des marqueurs selon l'importance des tweets correspondants (en fonction de leurs retweets, ajout aux favoris...) ou en fonction de dates ou mots clés.

Après avoir expliqué les différents modules disponibles sur notre petit logiciel et de quelle manière il est possible de s'en servir, nous allons utiliser celui-ci pour faire une analyse simple afin de montrer concrètement comment il est possible de s'en servir.

## II – Exemple d'analyse en utilisant uniquement le logiciel

### A) Récupération des données du compte @MasterSDIN



The screenshot shows a Java Swing window titled "tk (Ne répond pas)" with standard window controls. The main content area is titled "Récupération de données Twitter". It contains five text input fields with the following labels and values:

Label	Value
Consumer Key	IApI4jM5ykvZ4HCv4
Consumer Secret	HZoWIAunQl6FMDjq
Acess Token	I4w6CIncXIFE0WnRh
Acess Token Secret	41KtWNUOJwnZmW
Compte Twitter :	@MasterSDIN

Below the input fields are two buttons: "Retour" and "Valider". At the bottom of the window, there is a warning message in French:

Attention, en cliquant sur Valider, vous lancez le processus de récupération des données. Ce processus peut prendre un certain temps en fonction de la taille de la base de donnée. Vous pouvez suivre l'avancée de cette récupération via la console. Une fenêtre vous indiquera quand le processus sera terminé.

Il est possible de suivre l'avancement de la récupération de données twitter du compte spécifié via la console. Il est indiqué le nombre de comptes twitter récupérés, par rapport au nombre total de comptes à récupérer. Il s'agit ici des données de comptes followers du



compte étudié. Le compte MasterSDIN a 469 abonnés.

Lorsque la ligne “Rate limit reached. Sleeping for : 880” apparaît, cela veut dire que Twitter bloque la récupération de données pendant X secondes, ici 880. Cela peut donc prendre du temps pour récupérer les données, en fonction du bridage effectué par l’API Tweepy.

A noter qu’il est possible d’effectuer une récupération de donnée d’un compte twitter en plusieurs fois.

```
10 comptes twitter récupérés, sur 469
{'limit': 15, 'remaining': 4, 'reset': 1548676169}
{'limit': 15, 'remaining': 5, 'reset': 1548676170}
11 comptes twitter récupérés, sur 469
{'limit': 15, 'remaining': 3, 'reset': 1548676169}
{'limit': 15, 'remaining': 4, 'reset': 1548676170}
12 comptes twitter récupérés, sur 469
{'limit': 15, 'remaining': 2, 'reset': 1548676169}
{'limit': 15, 'remaining': 3, 'reset': 1548676170}
13 comptes twitter récupérés, sur 469
{'limit': 15, 'remaining': 1, 'reset': 1548676169}
{'limit': 15, 'remaining': 2, 'reset': 1548676170}
14 comptes twitter récupérés, sur 469
{'limit': 15, 'remaining': 0, 'reset': 1548676169}
{'limit': 15, 'remaining': 1, 'reset': 1548676170}
Rate limit reached. Sleeping for: 880
```

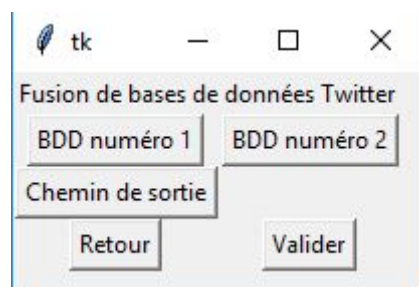
Le fichier correspondant à la base de données enregistrée, se trouvera dans le dossier @comptetwitter/Datas\_globales\_@comptetwitter. Ce dossier se trouvera à l’endroit où sont présents les différents fichiers Python permettant le lancement du logiciel.

 @MasterSDIN -->  Datas\_globales\_@MasterSDIN

Il est possible, de la même manière que pour le compte twitter du master SDIN, de récupérer les bases de données d’autres comptes twitter, afin de les fusionner par la suite. Ici par exemple, pour avoir une meilleure représentation du réseau lié au master SDIN sur Twitter, il pourrait être intéressant de fusionner la base de données du compte officiel du master SDIN, et celle du compte twitter du Master 2 des SDIN ou de certains représentants/professeurs du Master.

## B) Fusionner des bases de données Twitter

Comme nous l'avons dit précédemment, il est souvent utile de fusionner des bases de données récupérées sur Twitter à partir de certains comptes, pour obtenir une base finale plus grande, plus globale, et donc plus représentative et plus précise. Seulement, plus les bases de données récupérées seront nombreuses et conséquentes, plus cela prendra du temps à récupérer ces données. Il faut donc bien arbitrer entre le besoin d'augmenter la taille de la base de données Twitter pour augmenter la pertinence de l'analyse (mieux cerner un domaine) et le temps que cela va prendre.



En cliquant sur les boutons 'BDD numéro 1' et 'BDD numéro 2', il s'ouvre une fenêtre permettant à l'utilisateur de renseigner les fichiers voulus correspondant aux bases de données twitter qu'il veut fusionner.

Le bouton "Chemin de sortie" ouvre une fenêtre à l'utilisateur, lui permettant de renseigner l'endroit dans lequel il voudra que sa nouvelle base de données fusionnée soit enregistrée.

## C) Filter une base de donnée Twitter

Une fois la base de données constitué (fusionné ou non), il est possible d'appliquer des filtres afin de supprimer les comptes twitter qui ne paraissent pas pertinent pour l'analyse. Dans notre exemple, la base de données contient trop de comptes pour pouvoir la représenter graphiquement. Nous allons donc faire ressortir un nombre raisonnable de comptes, ceux que nous considérons comme les plus importants, en supprimant les

comptes twitter ayant moins d'un certain nombre de followers, amis et tweets. Nous allons également utiliser les filtres par mesure pour supprimer les comptes non pertinents pour faire une analyse d'importance. Dans cet exemple nous avons choisi d'utiliser la mesure "Closeness" qui traduit l'éloignement moyen aux nœuds du réseau.

La suite de l'exemple utilisera une base de données moins conséquente, par manque de temps.

tk

Filtre de bases de données Twitter

Base de donnée à filtrer

Filtrage par nombre de Tweets : ☒

Nombre de Tweets : 20

Filtrage par nombre de Friends : ☒

Nombre de Friends : 20

Filtrage par nombre de Followers : ☒

Nombre de Followers : 20

Filtrage par importance du compte (en terme de mesure) ☒

Choix de la mesure :

- Degré de centralité
- Closeness**
- Eigen Vector

Pourcentage des comptes les moins importants supprimés : 0.5

Ecriture de la base en fichier CSV : ☒

Chemin de sortie

Retour Valider

Il est également possible de suivre les filtres appliqués sur la base de données Twitter à travers la console.

```
Suppression de Elise car il a 4 followers (abonnés)
Suppression de Fanny car il a 17 followers (abonnés)
```

La nouvelle base de donnée filtrée sera enregistrée à l'endroit spécifié par l'utilisateur, en cliquant sur le bouton "chemin de sortie". Le fichier CSV correspondant sera présent à la même adresse.



	A	B	C	D	E	F
1	Screen name	ID	nb followers	nb friends	nb tweets	probleme
2	DugueAlan	328070080	61	166	202	non
3	DoryanGql	1286386724	132	122	3216	non
4	ClaraDaniel3	549468159	145	82	2576	non
5	diguiz_	1140844770	47	58	184	non
6	Romain_Iz	1506851334	159	134	1537	non
7	youneslass	1507269840	21	48	24	non
8	BigPAPOUCHE	509120852	63	110	1724	non
9	TheoNgr	710645997	244	237	6016	non
10	Oludolt	624162598	67	249	1183	non
11	LLrdx	1329581858	54	55	1228	non
12	KylianLeFlohic	163652840	235	153	2934	non
13	ClrtLea	2260283213	29	95	99	non
14	Ouzzouh	609005693	118	131	2910	non
15	MacheTeLAM69	1320863712	62	459	3168	non
16	CoxCollet	1325064186	187	249	3790	non
17	MohaRozay	391546793	110	213	3241	non
18	LeslieCorbel	501693686	1177	528	6173	non
19	Heloisebsrn	1063319191	171	131	3230	non
20	MaelHedin	1052435144	144	238	3222	non
21	jullivr	2332665669	45	58	115	non
22	charlottegql	1704650468	87	112	877	non

## D) Représentation de bases de données Twitter.

Une fois la base de données Twitter voulue filtrée, il est intéressant de représenter ces données et ces réseaux, afin de pouvoir procéder à une analyse visuelle de ce réseau. Nous allons donc utiliser une base de données filtrée auparavant afin de la représenter sous forme de réseau graphique. Il est proposé différents paramètres que l'utilisateur peut changer pour personnaliser la représentation de son réseau. Ce sont principalement des paramètres qui permettent une meilleure lecture du réseau.



tk

Représentation graphique de réseaux Twitter

Base de donnée pour construire le réseau

Nombre de graphe créés : 3

Titre du graphe : Titre Reseau

Zoom sur le graphe : 40

Activer les axes : ☐

Couleur des liens : black

Taille des liens : 5

Couleur des noeuds : black

Taille des noeuds : 4000

Afficher des labels : ☒

Type de labels : screen\_name

Tailles des labels : 4

Couleur des labels : black

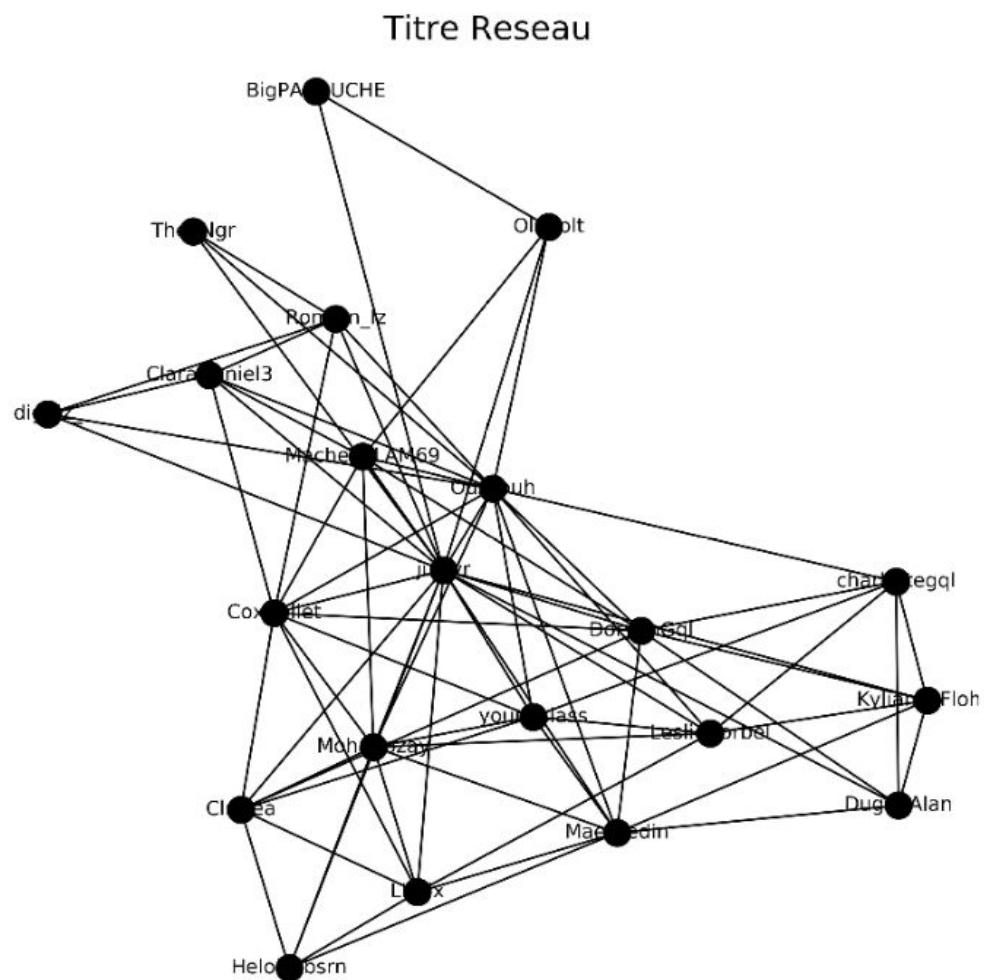
Dispersion des noeuds\* : 0.7

Chemin de sortie : Browse

Format de sortie : png

Retour Valider

Le réseau pourra être visualisé et analysé à partir de l'adresse renseigné par l'utilisateur.



## E) Récupération de tweets

Il est également possible d'analyser les dynamiques de tweets de certains comptes identifiés. Pour cet exemple, nous allons analyser la fréquence de tweets de la part du compte Master SDIN sur le sujet du Web We Can, et de la cybersécurité. Pour cela il faut tout d'abord récupérer la liste des tweets du compte twitter voulu.

Récupération de Tweets

Consumer Key: lApl4jM5ykvZ4HCv4

Consumer Secret: HZoWIAunQl6FMDjq

Access Token: l4w6CIncXIFE0WnRh

Access Token Secret: 41KtWNUOJwnZmW

Compte Twitter : @MasterSDIN

Chemin de sortie

Retour Valider

Attention, en cliquant sur Valider, vous lancez le processus de récupération des Tweets.  
Ce processus peut prendre un certain temps en fonction de la taille de la base de donnée.  
Vous pouvez suivre l'avancée de cette récupération via la console.  
Une fenêtre vous indiquera quand le processus sera terminé.

Il est possible de suivre l'avancée de cette récupération sur la console, de la même manière que pour la récupération de données de comptes twitter.

```
getting tweets before 691964403676246015
...400 tweets downloaded so far
getting tweets before 527835144645001216
...600 tweets downloaded so far
getting tweets before 344108198170595328
...799 tweets downloaded so far
getting tweets before 38349030564171775
...885 tweets downloaded so far
getting tweets before 4545991176880127
...885 tweets downloaded so far
```

## F) Filtres de tweets

Une fois les tweets récupérés, il est souvent intéressant de les filtrer pour faire ressortir uniquement les tweets qui semblent pertinents à analyser en fonction de ce que l'on veut montrer.

Pour cet exemple, ce qui nous intéresse sont les tweets sur l'évènement Web We Can (tweets et retweets) et sur la cybersécurité.

Suivi sur console :

```
RT @tpenard: Demandez le programme du @WebWeCanEcotic J-10 #Rennes http://t.co/WHBqlubqZu cc @FondationR1 @UnivRennes1
RT @LeMagNumerique: Le Web We Can, digital day par @Master2Ecotic http://t.co/rnBokhsaxo Qui fête ses 10 ans pour l'occ
sion. #Rennes
RT @pareto35: You,digital explorer,le 21mars c'est le @WebWeCanEcotic digital festival.On fête les 10ans @Master2Ecotic
http://t.co/8W8VUR3...
RT @WebWeCanEcotic: Nous avons la joie de vous annoncer la labellisation de l'événement #WebWeCan FrenchTech !
RT @WebWeCanEcotic: L'événement Web We Can aura lieu le 21 mars 2015, 7 Place Hoche à Rennes.
Plus d'informations très prochainement. Suive...
RT @secteur_sud: Recherche stagiaire Redacteur-Animateur Web pour http://t.co/Ca7IEHs2Rs à Rennes. Fiche de poste sur d
mande. RT plz
NOMBRE DE TWEETS AVEC AU MOINS UN DES MOTS CLES : 29
```

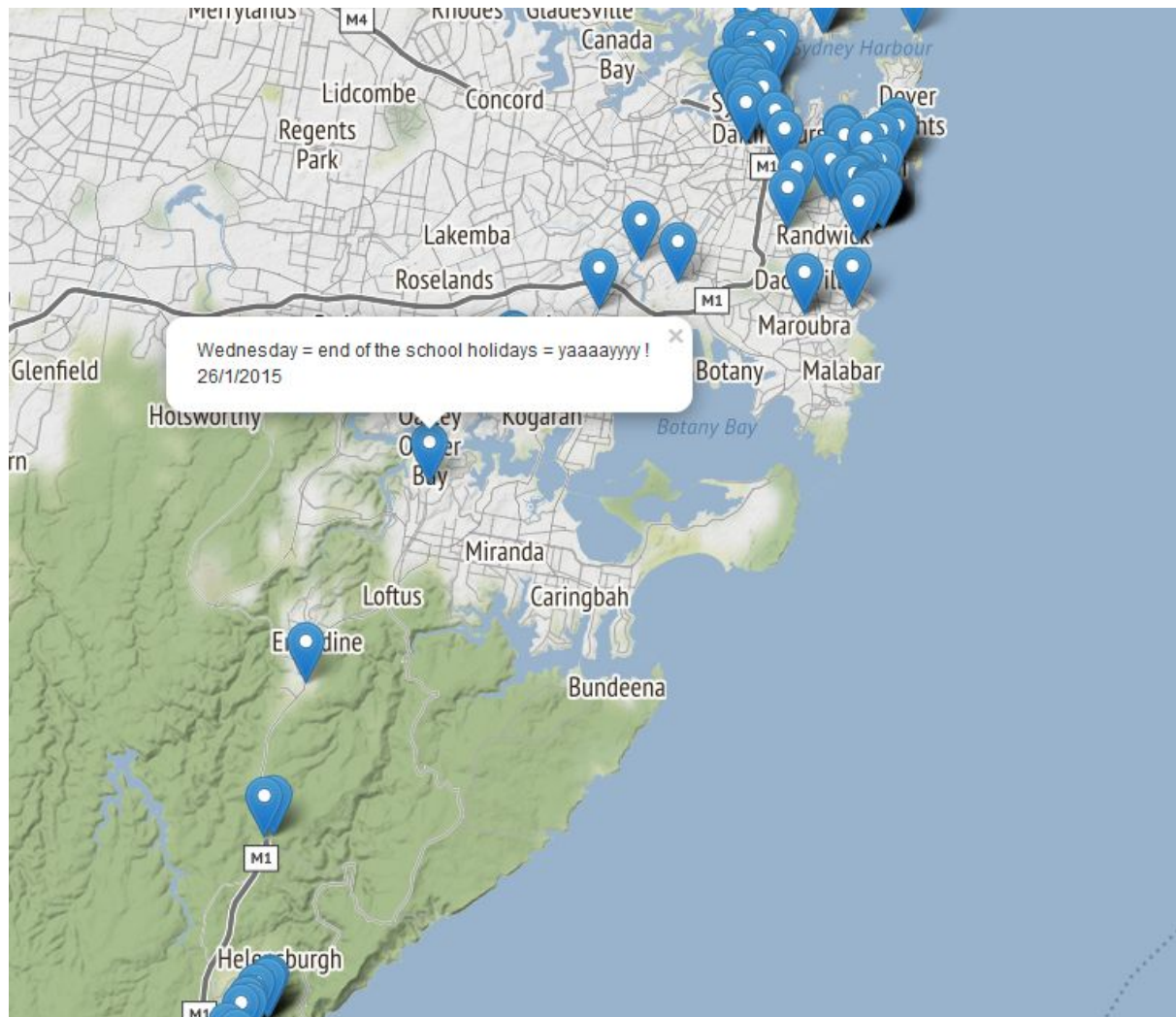
## G) Représentation de tweets

Une fois la base de données de tweets filtrée pour ne garder que les tweets qui semblent intéressants à analyser, il est possible de représenter ces tweets de différentes manières : soit sous forme de tableur (qui permet d'utiliser ensuite tous les outils disponibles à partir du tableur), soit sous forme de carte géographique (représentation des tweets géolocalisés).

Sous forme de tableur :

Screen name	tweeté le :	Nombre de likes	Nombre de retweets	Texte
MasterSDIN	2019-01-25 10:08:13	0	7	b"RT @cyberguerre: J'a le plaisir d'intervenir ce vendredi \xc3\xa00 @EcoRennes1 @UnivRennes1 pour parler #cybers
MasterSDIN	2019-01-25 08:47:40	0	10	b"RT @ericdarmon: Web We Can 4 @MasterSDIN : c est aujourd'hui \nSensibilisation sur la cybers\xcc3\xa9curit\xcc
MasterSDIN	2019-01-22 14:53:57	0	9	b"RT @UnivRennes1: D'ici \xc3\xa00, rdv au #WebWeCan vendredi 25/01 @EcoRennes1\nUn \xc3\xa09vxc3\xa9neme
MasterSDIN	2019-01-15 15:02:37	0	7	bRT @BayolSteve: Web We Can 4 by #MasterSDIN : Sensibilisation sur la cybers\xcc3\xa9curit\xcc3\xa9 avec des exp
MasterSDIN	2019-01-03 10:08:13	22	23	bWeb We Can 4 by #MasterSDIN : Sensibilisation sur la cybers\xcc3\xa9curit\xcc3\xa9 avec des experts et des hacke

Sous forme de carte géographique (les données utilisées sont issu d'un compte twitter ayant tweeté de manière géolocalisé) :



# Conclusion

Pour conclure, il est intéressant de signaler que ce petit logiciel a été réalisé en programmation objet avec de nombreuses fonctions, ce qui permet un développement continu de nouveaux outils, qu'il est possible d'ajouter au logiciel. Les modules sont indépendants les uns des autres, et n'importe qui peut ajouter des fonctionnalités ou des modules, sans compromettre le fonctionnement des autres modules ou fonctionnalités.

Ce petit logiciel permet donc à tous de constituer ses propres réseaux Twitter et d'en effectuer une analyse simple. Il faudra cependant prendre en main le langage de programmation Python si l'utilisateur veut effectuer des analyses plus poussées et plus personnalisées de dynamiques sur twitter.