Sommaire :

❖ Description du sujet

Le projet consiste à créer un robot capable d'extraire des données à partir de Twitter, des données telles que les « followers », « following », « biographie » ainsi que les tweets d'une personne souhaitée.

Dans ce projet nous pouvons distinguer 4 missions :

- Extraction des données
- Conversion des données
- Traitement des données
- Création de l'interface utilisateur

1) L'extraction de donnée

Pour extraire les données d'un compte Twitter il nous faudra préalablement avoir le nom du profil sur lequel on voudrait acquérir les données.

Le robot devra extraire des données précises de l'utilisateur :

- ses "followers"
- ses "following"
- ses "tweets"
- son pseudonyme
- sa biographie

Il devra paraître humain pour éviter la détection de bot faite par twitter automatiquement.

2) La conversion

Les données que nous avons extraites du profil Twitter par le robot ne seront pas directement dans le bon format (probablement de type HTML ou XML). Ces données devront être transformées et adaptées pour pouvoir être rentrées dans une base de données préalablement configurée pour les accueillir. Nous devrons décider du meilleur modèle relationnel possible pour gérer ces données dans notre base de données.

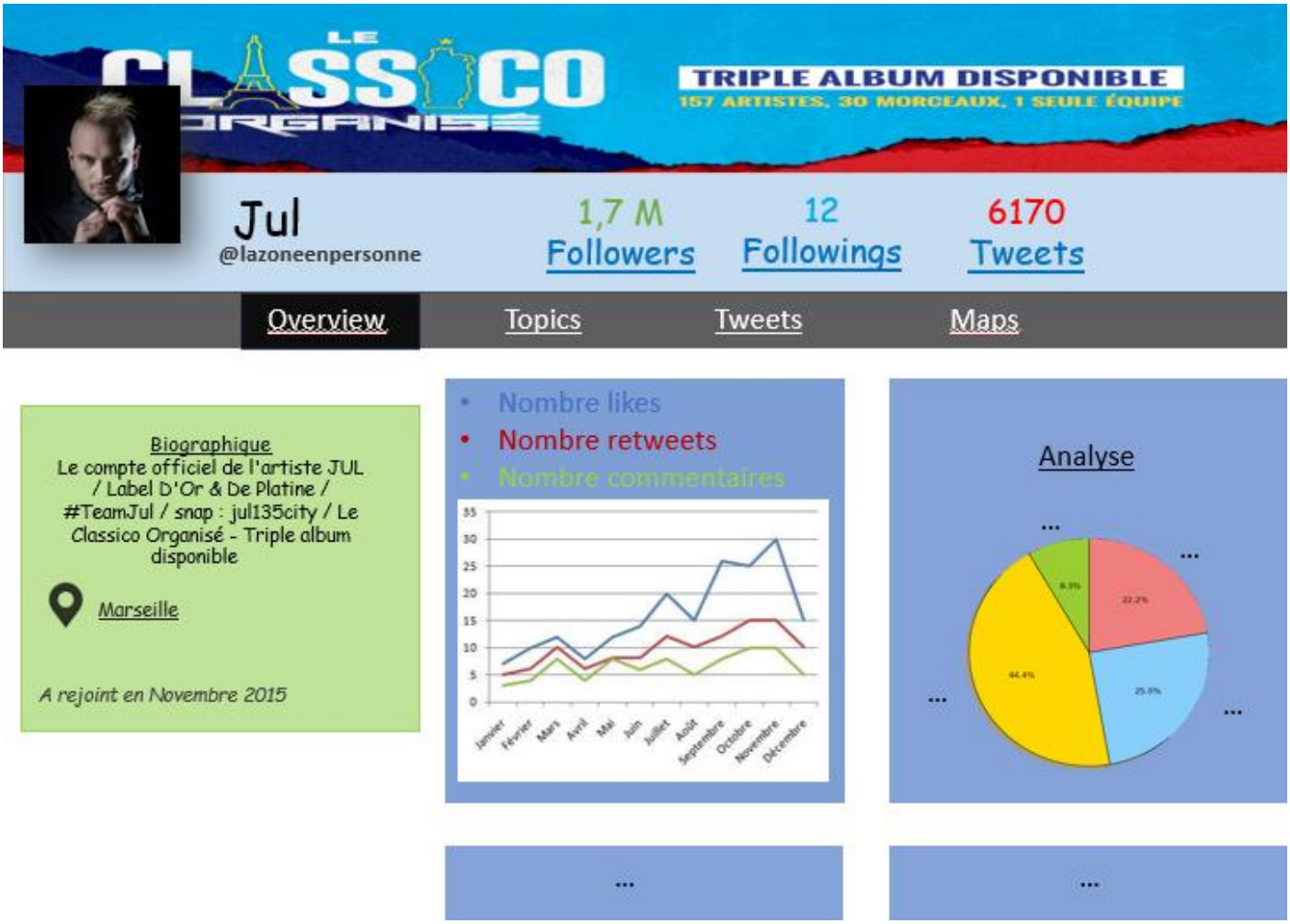
3) Le traitement

Après avoir été stockées dans notre BDD, ces données devront être exploitées par différentes requêtes demandées par l'utilisateur de l'application par exemple. La base de données est remplie des informations du profil Twitter, il nous faudra tout simplement exécuter les requêtes pour pouvoir les visualiser.

4) La création d'une interface

La partie "finale" du projet sera la création d'une Interface Homme Machine permettant à n'importe quel utilisateur novice de demander à la base de données des requêtes et recevoir des infos simplement en cliquant sur des boutons.

L'application demandera le nom du profil sur lequel l'utilisateur souhaite récupérer les données et affichera ensuite un résumé de ses données mais pourra également filtrer son contenu ou voir ses données sous un angle particulier. L'interface sera le plus possible "user-friendly".



Sur cette maquette, on aurait plusieurs parties, avec la partie de base « Overview » montrant les informations basiques du compte, avec biographie, followers, followings et nombre de tweets mais aussi quelques statistiques et graphiques en plus. On pourrait faire en sorte de cliquer sur « Followers », Followings » et Tweets pour voir une liste de ces derniers.

Il pourra y avoir d'autres parties comme « Topics » montrant des informations sur les sujets appréciés du compte et certaines statistiques autour de cela.

Enfin nous pourrions avoir une partie « Maps » avec un affichage très visuel de relations entre des comptes par exemple.

❖ Etude de projets similaires existants

Concernant les projets similaires existants, il en existe énormément mais sont souvent payants ou incomplets. En effet, la plupart des applications ou sites web permettant l'obtention et l'analyse de données twitter sont fondées sur un besoin d'information en particulier.

Certains sont plus centrées sur des aides factuelles donnant des chiffres précis sur des informations tels que le "retweet" ou le nombre de "like" par tweet, alors que d'autres tentent des approches plus globales avec des relations entre différents compte twitter pour, dans de nombreux cas, combien de personnes sont touchées par un tweet par exemple.

Certains proposent des fonctionnalités que d'autres ne proposent pas. Ce que nous voudrions faire, c'est une interface qui regroupe plusieurs fonctionnalités afin de répondre aux plus de besoins possibles, et ces choix peuvent commencer en regardant les projets déjà existants.

Nos recherches se sont donc faites autour de sites web (pour la plupart des solutions trouvés) et utilisant différentes technologies.

Malheureusement nous n'avons pas pu avoir connaissance des technologies utilisées dans les cas étudiés car ils sont souvent discrets mais certains temps de latence feraient deviner qu'il n'a y pas de base de données associés mais seulement des recherches directes sur le site Twitter (le site et ses données étant utilisés comme une base de données à part entière). Cependant nous ne pouvons être sûrs de rien et tout ceci n'est que des suppositions.

Nous avons choisi de les montrer dans l'ordre de présentation suivant :

-Twitter Analytics

-Twitonomy

-MentionMapp

-Foller.me

-followerWonk

Chacun des projets présentés a été choisi car il est particulier et pourrait apporter un nouvel aspect intéressant à notre projet. Une page présente ses caractéristiques et un screenshot de son interface.

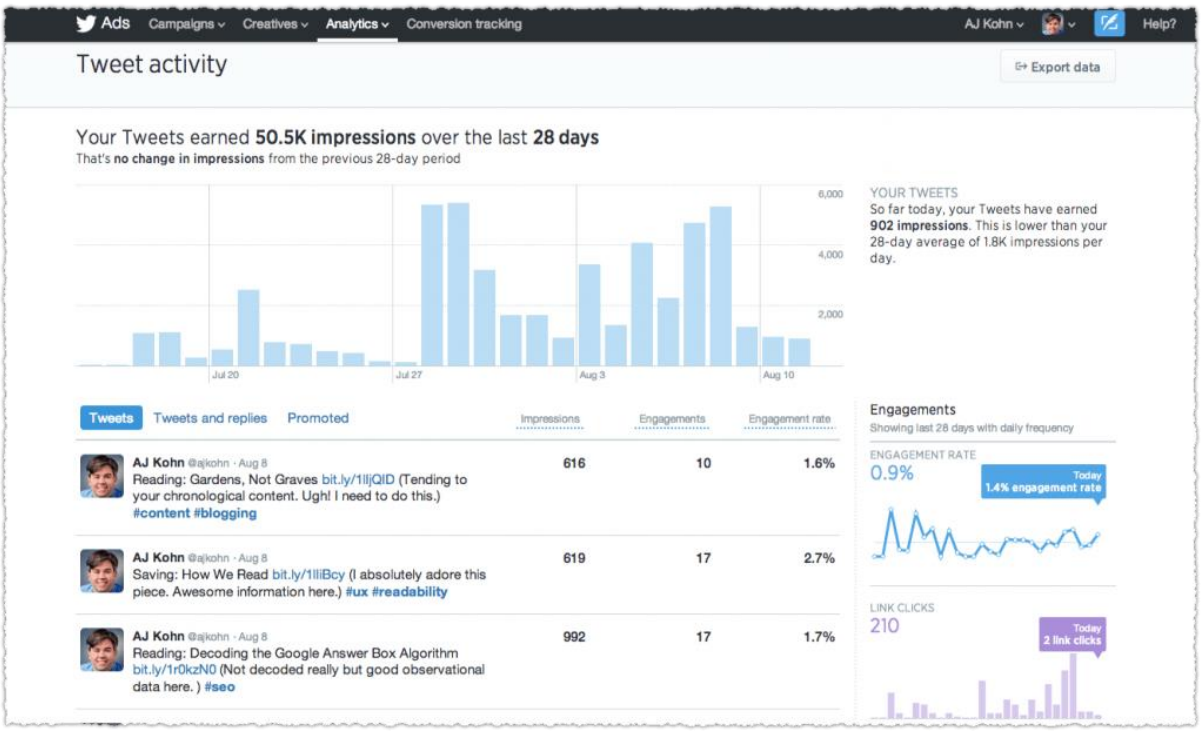
Les projets connu et intéressant se rapprochant du nôtre sont les suivants :

Twitter Analytics

Tout d'abord TwitterAnalytics, un site web créé par Twitter directement pour que les utilisateurs intéressés n'aillent plus sur des sites tiers. Ce site est donc probablement celui ayant la plus grande fiabilité dans ses informations dû à son lien plus que direct avec twitter même.

Un site très intéressant est assez graphique qui affiche sous forme de courbe votre nombre d'interactions et de réactions sur tout twitter ce dernier mois (28 jours). Le site est bien fait, complètement gratuit, intuitif et donne beaucoup de données. Il compare même avec les mois antérieurs pour savoir s'il y a plus ou moins d'interactions comparées au passé. C'est un bon outil pour un particulier intéressé par ses statistiques, mais il est vrai que comparé aux autres sites que nous allons voir, Twitter Analytics n'apporte peut-être pas tant de nouvelles informations que cela.

Le problème notable du site étant que nous ne pouvons qu'étudier les statistiques de notre propre compte.



Twitonomy

Site web et application portable, Twitonomy permet beaucoup de choses et même plus que Twitter Analytics. Les plus gros points forts de ce site sont les graphiques, plus nombreux et plus paramétrables que ceux de Twitter Analytics, ils permettent de bien pouvoir analyser les données autour du compte d'une entité. Je pense que ce site est un excellent outil professionnel pour une entreprise lorsqu'il s'agit d'étude de marché ou d'étude de communication autour du twitter de l'entreprise ou de concurrents. En effet la plus grosse différence entre Twitter Analytics et ses concurrents se fait sur l'étude d'autres comptes que le siens : on peut "espionner" ou en tout cas analyser les statistiques que produit un compte en particulier pour, par exemple, connaitre le nombre de personnes différentes ayant vu son tweet ; pour ensuite en déduire le prix d'une publicité sur ce dernier par exemple.

Un gros point de divergence avec les autres outils présent est que ce Twitonomy n'est pas 100% gratuit et que l'on doive payer pour avoir accès à tous les « features » possibles

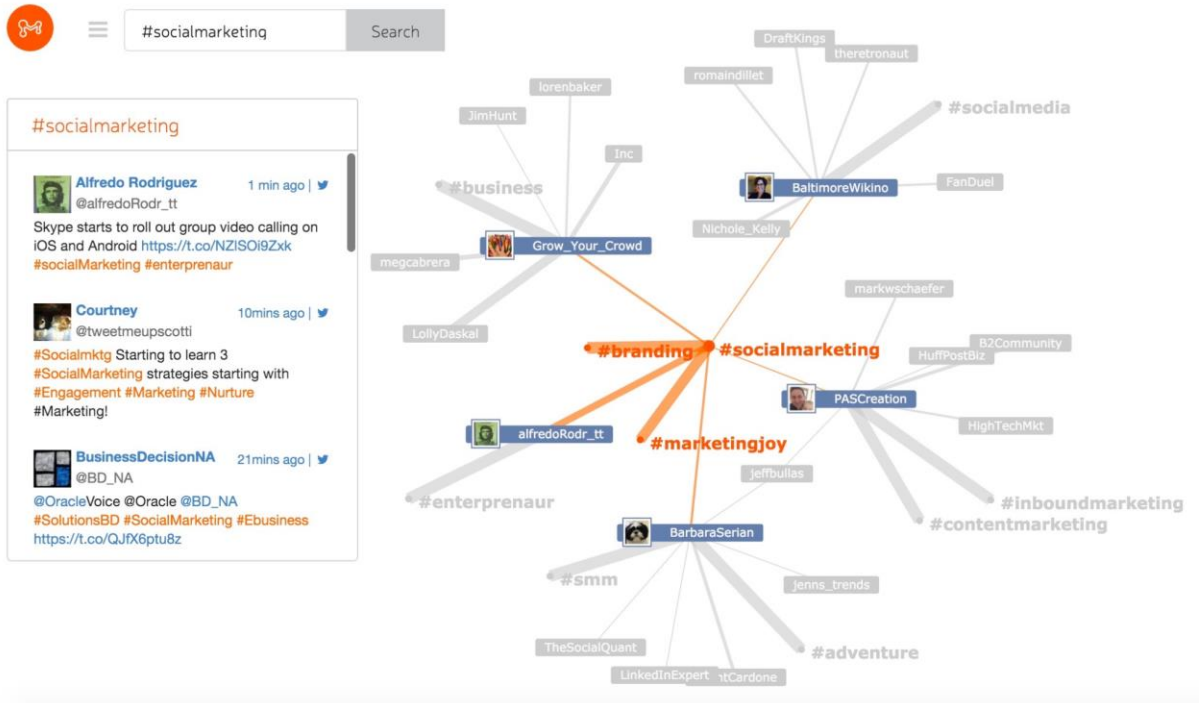
En somme un très bon outil, conseillé pour des entreprises pour son nombre d'informations et sa complexité (sans oublier que sa version complète est payante).

Mentionmapp

De tous les sites ou applications d'analyse et d'extraction de données de Twitter, MentionMapp est le plus inventif. Il permet l'affichage de données simples d'accès sur d'autres solutions du même genre déjà cité, mais a une option principale - faisant tout l'attrait du site - étant un graphique ou plutôt une carte de tous les utilisateurs (symbolisé par leurs photos de profils sur twitter) ayant interagi à vos tweets de près ou de loin reliés entre eux en fonction de l'interaction.

Cela est extrêmement visuel et peut permettre beaucoup de choses. On peut mieux visualiser qui s'intéresse à nous, qui est possiblement intéressé par nous ou bien même pour des comptes étant suivi par énormément de gens de se rendre compte de l'impact qu'ils ont. Cela peut aussi avoir un attrait dans le cadre d'une entreprise pour étudier l'environnement en interaction avec une entreprise concurrente par exemple.

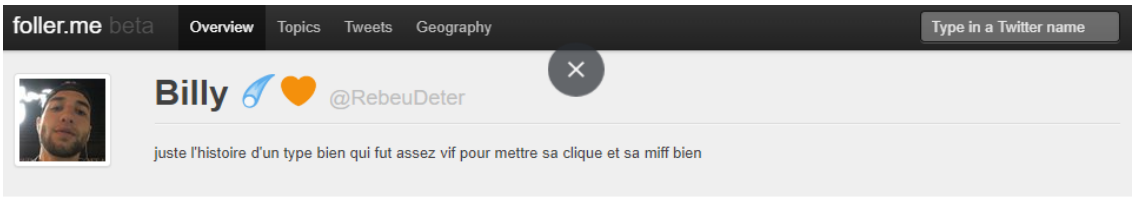
Cet outil très visuel est quasi complètement payant mais reste très intéressant et très actuel. De tous les projets étudiés, Mentionmapp est celui qui se départage le plus par sa différence et sa touche visuelle.

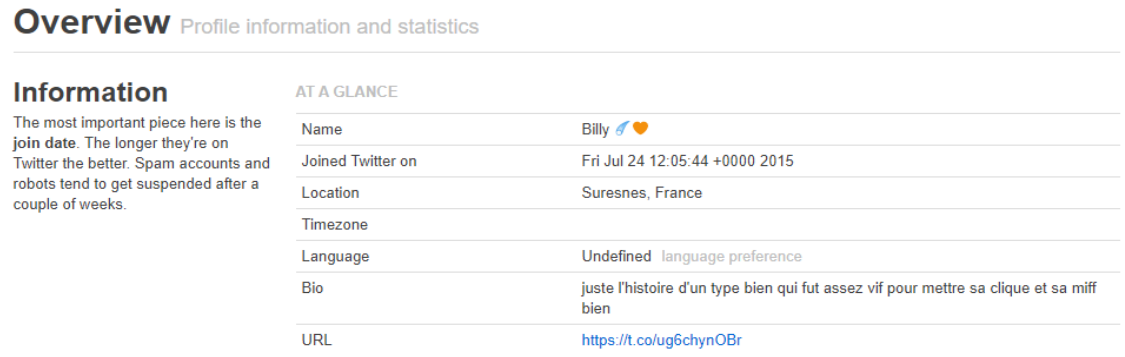


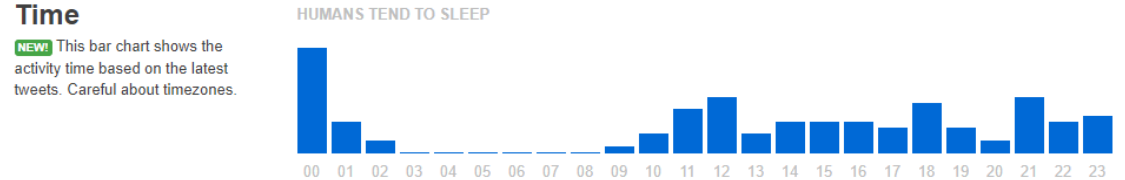
Foller.me

Site web très simple qui est plutôt intéressant pour une étude simple sur un utilisateur en particulier. Plus de l'ordre de l'intérêt que de l'analyse, ce site reste intéressant pour son étude des "Topics". Après avoir tapé le nom de l'utilisateur à étudier, le site prend un temps assez lents comparé aux autres pour étudier le profil en question (d'ailleurs le site explique qu'ils utilisent directement l'API de Twitter pour l'analyse des profils, je suppose donc qu'ils n'utilisent pas de base de données mais font l'extraction des données seulement à la demande de l'utilisateur. Puis affiche des données basiques telles que le nombre de followers, la date de création du compte ou son nombre de tweet au total. Mais La catégorie Topics est plus novatrice et affiche les Hashtag les plus utilisés, les mots les plus cités par le compte dans tous ces tweets, en plus d'une image des personnes les plus mentionnées par le compte en question.

D'autres outils y sont présents, assez accessoires mais toujours intéressant tel que le nombre de tweets par heure, pouvant donc montrer - pour les comptes les plus actifs - les heures de sommeils de la personne possédant le compte. Il y a aussi un outil comptant le nombre de réponses à tweets pour 100 tweets ou encore le nombre de tweets avec des liens dedans et si oui montrer les plus utilisés.





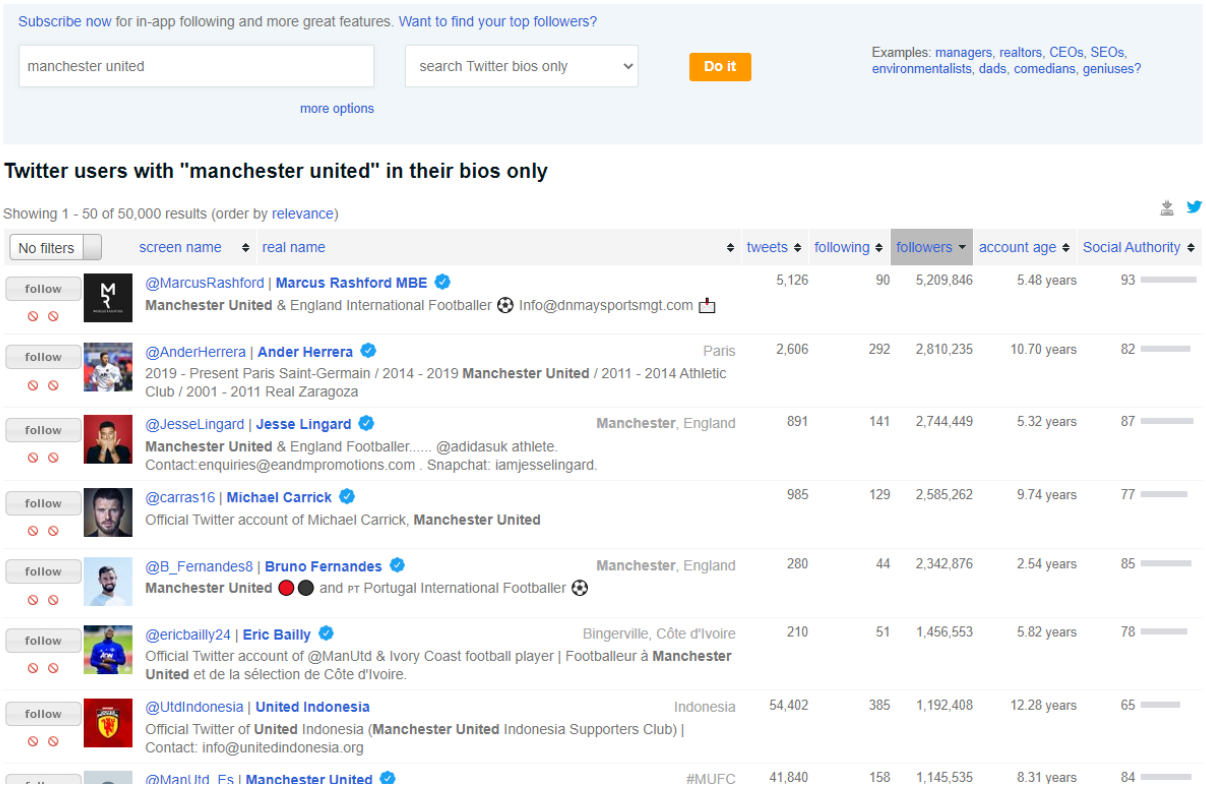


followerWonk

Ce site web gratuit ("freemium", ce qui signifie qu'il a des fonctionnalités gratuites mais peut en obtenir d'autres en payant) à la grande caractéristique de pouvoir chercher tous les comptes possédants certains mots dans leurs biographies twitter.

Vous n'avez qu'à taper dans la barre de recherche les mots que vous voulez et ils vous affichent tous les comptes ayant ces mots (précisément) dans leur biographie, vous pouvez triez la liste des comptes par nombres de followers, l'âge des comptes (leurs dates de création), leurs nombres de tweets ou bien leurs nombres de following.

Les comptes trouvés peuvent enfin être triés par "Social authority" une variable créée par le site donnée à chaque utilisateur en fonction de vous, permettant normalement d'optimiser vos recherches. C'est une échelle de 1 à 100 où 100 serait le plus influent des comptes et cela en fonction de vous, selon un calcul gardé par le site mais quand même vaguement expliqué ainsi : la variable de "Social Authority" est composé du taux de retweet sur les 100 derniers tweets du compte en question, de la récence des tweets en question le tout basé aussi sur les retweets du profil utilisateur.



La liste ci-dessus nous montre la diversité des projets recensés capable d'extraire des données twitter, chacun ayant sa spécificité. De leurs interfaces, à moyen de stocker les données, chacun des projets présentés peut aider notre étude du sujet.

Notre solution regroupera des fonctionnalités de ces applications afin de pouvoir proposer à l'utilisateur de l'application une variété de fonctionnalités/options le plus vaste possible.

❖ Etude technique

Pour l'instant, nos choix de technologie ne sont pas encore faits, mais voici les technologies utilisables pour réaliser ce projet :

Base de Donnée

Pour les bases de données, le choix évident serait forcément d'utiliser Oracle dataBase. En effet c'est avec cela que nous travaillons depuis 1 an et demi pour tous nos projets et nous connaissons bien le langage SQL et PL/SQL dont nous aurons besoin pour gérer les tables et les requêtes.
De plus, c'est cette base de données qui est utilisée par l'IUT et nous pourrions utiliser son serveur directement.

image : oracle SQL Developer

SQL Developer sera le logiciel utilisé pour la création des tables.

- L'extraction des données

Pour cette partie, le langage choisi sera probablement java puisque c'est le langage avec lequel nous avons le plus d'expérience, cependant les API que nous allons possiblement utiliser sont toutes présentes sur beaucoup d'autres langages très populaire tel que le C#, javascript ou python.
Pour l'extraction nous avons principalement 2 choix :

-API Twitter

C'est l'API créé par Twitter eux-mêmes, donnant accès à quelques fonctionnalités de leur application.
Cependant cette API, bien qu'être sûr, de par son lien direct avec Twitter, n'est pas très utile puisqu'elle n'offre que très peu de possibilités, et ce, surtout au niveau de la récupération de données.
L'interface de programmation d'application de Twitter ne permet que - après avoir dû se connecter à un compte twitter déjà existant - de retweeter, éditer des tweets, rechercher des tweets sur Twitter directement et simplement créer des tweets.

On peut donc facilement comprendre que nous allons être vite bloqué par un manque de fonctionnalité de l'API.

-API Selenium

Cette API très connue permet énormément de choses par rapport aux sites web et aux navigateurs en général.

- Travail sur la base

Pour le traitement des données et tout ce qui est des requêtes autour des tables, nous utiliserons probablement Java avec l'API JDBC déjà bien connu de nous tous. Avec JDBC nous pouvons nous connecter et demander des requêtes aux tables créées sur SQL Developer. Nous pouvons d'ailleurs créer ces mêmes tables sur JDBC Directement même si l'affichage sera peut-être moins graphique qu'avec d'autres outils plus centrés sur les bases de données.

Interface Utilisateur

Pour l'interface Utilisateur, 2 choix s'offres à nous :

- Java, une application en java complet grâce à Java Swing qu'un utilisateur devra installer sur son ordinateur pour utiliser. Le gros point fort de cette solution est que tout le projet serait en java ce qui réduirait la chance de problèmes de compréhension entre les différentes parties. Nous pouvons aussi noter l'existence d'une nouvelle partie implémentée à java depuis peu : le javaFX. Nous n'avons jamais rien vu sur ce nouveau compartiment graphique de java mais cela peut aussi être une option valable à ne pas écarter.

- HTML, CSS et javascript, un site web.
En effet, comme on a pu le voir dans l'étude des projets similaires, presque tous les projets existants ont été fait sous la forme d'un site web ou presque (pour les applications mobiles, elles sont faites en java).

Même si c'est 2 choix semblent bons, nous pensons partir sur une application en java puisque cela n'est pas majoritaire et peut donc être intéressant, mais aussi puisque le projet porte surtout à se concentrer sur la java. L'application en java Swing sera donc la voie sur laquelle nous avons le plus de conviction.

❖ Problèmes possibles

1) Le premier problème que nous allons rencontrer est le système de bannissement de bot que Twitter et d'autre sites web ont mis en place depuis un certain temps.

Nous devrons veiller à ce que le Bot se comporte de manière humaine c'est-à-dire qu'il navigue sur la page web avec un temps paramétré. Il ne faut pas qu'il enchaîne les actions de manière instantanée, sinon Twitter pourra conclure qu'ils ont affaire à un robot.

2) Le deuxième problème qu'il serait possible de rencontrer est au niveau du stockage de ces données. Si on doit récupérer les tweets d'une personne qui a tweeter plus de 12 000 fois il sera difficile de stocker ce grand nombre de données. La solution serait de mettre en place une date de laquelle on partirait jusqu'à une autre date afin de récupérer les tweets de ces périodes.

   Si les tweets sont tout de même supérieurs à une capacité donnée et supportable par la base de données, on enverrait une erreur ou alors l'opération de récupération des données se fera en plusieurs fois afin de récupérer une partie des tweets, les afficher au client, les supprimer en gardant le dernier tweet affiché, recommencer à partir du dernier tweet affiché et cela jusqu'à la fin de la période donnée en paramètre.

Pour illustrer la solution, voici un schéma :