



UNIVERSITÉ
DE LORRAINE

IUT nancy
Charlemagne

Projet Tutorée 2018-19

Sommaire :

- I. Problématique
- II. Description
- III. Analyse
- IV. Etude de l'existant
- V. Besoins
- VI. Solutions
 - i. Conception/Méthode
- VII. Sélénium
- VIII. Interface Homme Machine

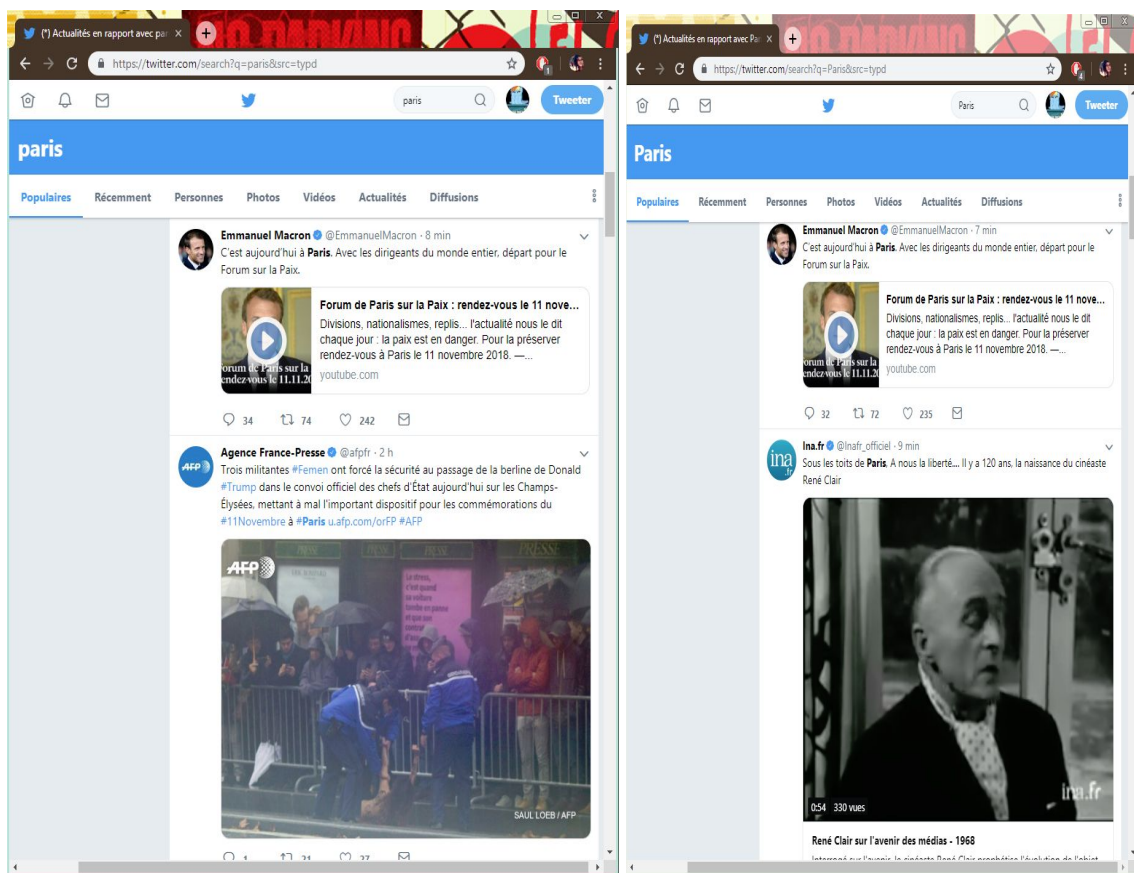
I. Problématique

Le système de recherche de Twitter n'est pas absolu, en effet la base de données limites les résultats.

Donc on peut faire plusieurs fois une même requête et obtenir des résultats différents.

Et les recherches complexes sont impossibles.

Exemple : La recherche du mot "paris"



II. Description

Le but de ce projet est de créer une application qui permettra de récupérer toutes les infos d'un compte twitter ciblé et les ajouter à une base de données, un des but est de pouvoir récupérer les infos d'un compte sans connexions internet mais elle permettra aussi de récupérer des infos sur les followers de la cible si demandée (si le compte est déjà renseigné dans la base de données).

III. Analyse

Nous avons plusieurs solutions qui s'offrent à nous pour répondre à cette problématique:

Utiliser l'API de Twitter, mais celui-ci nous contraint de :

- Requête par seconde : 10
- Requête par minute : 30
- Requête incomplète (filtrées par Twitter)
- La plupart des librairies sont pour Python

Ce qui n'est pas optimal pour l'usage que l'on fera. De plus, l'apprentissage d'un nouveau langage se présenterait comme un frein à l'avancement du projet.

Donc nous nous tournerons vers une base de données, et d'un robot nommé Sélénium, que nous décrirons l'utilité plus tard.

IV. Etude de l'existant

Il y a déjà des applications similaires :

Le moteur de recherche que Twitter nous propose mais comme dit plus tôt il est contraignant.

Le site Topsy qui permettait (car racheter puis supprimer par Apple) de trier les recherches, il utilisait l'API de twitter et puisqu'il était partenaire direct de Twitter, il n'était pas restreint par le nombre de requête.

Le site Twinitor qui permettait de réaliser des recherches Twitter diverses avec des filtres où afficher les résultats de façon chronologique, mais comme Topsy ce site n'existe plus.

V. Besoins

Ne pouvant pas utiliser l'API de twitter trop contraignante, il nous faut si on liste tous nos besoins : d'une base de données relationnelles, d'un compte Oracle, du robot Sélénium, de plusieurs comptes Twitter (que nous créerons, donc un total de 5 en comptant notre tuteur de projet), du langage Java (avec les librairies permettant de manipuler des base de données : JDBC).

VI. Solutions

La solution que nous avons choisis, qui est celle proposée par notre tuteur, est l'utilisation de l'outil Sélénium qui nous permettra de simuler les actions d'un humain, c'est-à-dire que après configuration il permettra d'ouvrir Twitter, de se connecter, de faire les recherche que l'utilisateur aura demandé et de stocker toute les infos qu'il aura pu récupérer sur la base de donnée prévue à cet effet.

Le problème avec cette solution est la pauvreté de notre base de données, en effet il est illégal de récupérer des données de comptes qui ne nous appartiennent pas, donc il faudra utiliser notre robot sur nos propres comptes. Mais on peut encore enrichir notre base en demandant l'autorisation de proches pour pouvoir utiliser leurs données avec notre robot.

Une question importante à se poser aussi sont le type de données que nous voulons récupérer, nous pouvons remarquer qu'il y a beaucoup d'informations à récupérer si nous voulons tout récupérer donc dans notre intérêt nous devons définir quelles données nous voulons récupérées et il faut faire attention à ne pas récupérer des données semblables qui peuvent apparaîtrent plusieurs fois.

Si nous prenons pas en compte ces détails notre base de données peut devenir non viable très vite. Il faut aussi définir des limites du nombre de données à récupérer sur un compte car si nous voulons alimenter notre base de données passant par compte ayant une très grande quantité de données (tweet, aime, liste de followers et follow) cela pourrait être très contraignant car nous ne disposons pas d'un espace stockage important et remplir notre base de données que de données provenant d'un seul compte est inutile.



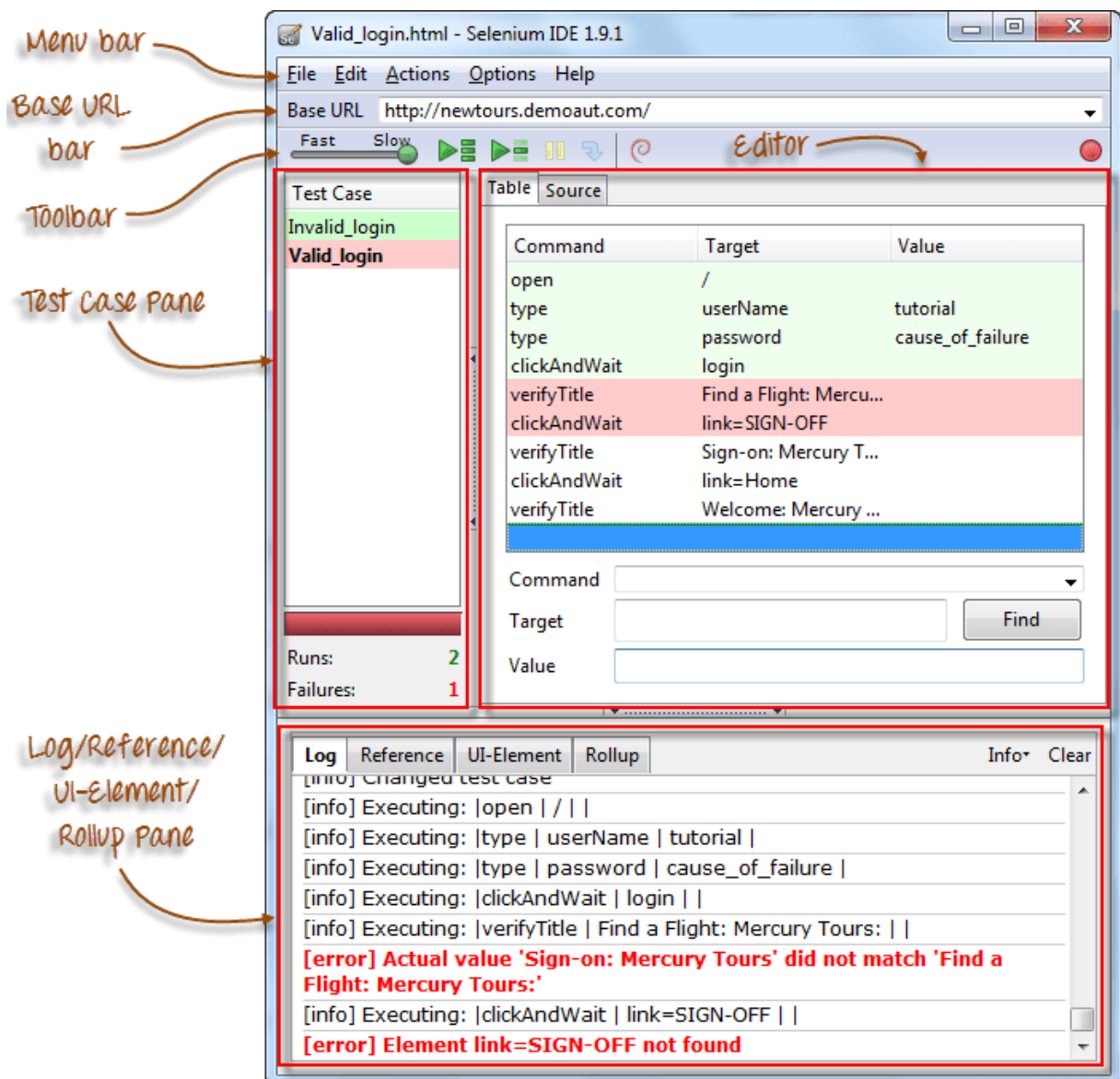
VII. Sélénium

Dans notre cas,

Selenium va récupérer les différentes données disponibles d'un profil, comme la bio (petite description du profil), les tweets, le nom d'utilisateur, le pseudo, les personnes qui suivent le profil, les personnes que suit le profil, la localisation, les mentions "j'aime" et les retweets. Plus tard on pourra définir un nombre de 'saut' à faire pour récupérer les données des abonnés du compte de départ.

C'est-à-dire que quand on analyse un profil, on regarde ses followers, et on analyse ensuite ses profils, et on limitera donc le nombre de saut à faire avant de s'arrêter, car cela croît exponentiellement et on peut se retrouver très rapidement avec 10 000 profils à analyser.





VIII. Interface Homme Machine

Nous pouvons développer une interface homme machine si nous avons le temps mais ce n'est pas une priorité. Mais si la création d'une IHM venait à aboutir elle consisterait en une page qui permettrait de faire les recherches sur la base de données et qui lancerait une recherche si la recherche n'existe pas sur la base de données. Voici à quoi ressemblerait l'ihm :

la page posséderait deux emplacements de texte, un pour écrire le mot clé à rechercher et l'autre pour afficher les résultats, d'un bouton pour lancer la recherche et d'une case à cocher qui permettra de lancer une recherche directement sur twitter sans passer par la base de donnée. On pourra ensuite voir plus loin et rajouter quand le programme sera en état de marche on un menu déroulant pour pouvoir affiner les recherche à une catégorie tel que les pseudo ou les descriptions des personnes.

liste des données d'un profil twitter :

@nomducompte (inchangeable)

nomducompte (changeable)

photo de profil

photo de bannière

biographie

localisation

site (ex youtube/facebook/blog ...) plusieurs possible

date de naissance

date d'inscription

abonnements -> autres profils

abonnés -> autres profils

j'aime -> tweet

liste(abonnés à :

-Abonnez-vous à une liste. Après ça, elle apparaîtra ici.

membre de :

-Il semblerait que personne ne vous ait encore ajouté à une liste. Peut-être un jour.)

moments(Les Moments sont des histoires organisées au sujet de ce qui se passe dans le monde — alimentées par des Tweets. C'est facile de créer votre propre histoire avec les Moments Twitter. Une fois faits, ils apparaîtront ici.)

tweets [

- tweets : liste de 'tous les tweets'(retweet inclus)

- tweets et réponses : tweet des conversation (il n'apparaissent pas dans 'tous les tweets') + les autres tweets (retweet inclus)
- média (liste de tous les tweets contenant un média vidéo,audio,images ...)

les tweets contient aussi potentiellement :

- des réponses avec d'autre tweet
- des retweets par d'autres profils twitter
- des j'aime par d'autres profils twitter
- un réponse par message privé si il est possible d'envoyer un message privé impossible quand le compte refuse les message privé des compte qu'il ne suit pas par exemple)

pour nos tweets on peut retrouver encore plus d'informations et statistique :

- impressions (nombre de vues de ce tweet)
- engagements totaux (nombre de d'interactions sur le tweet)
- clics sur le lien (nombre de clique sur le lien si le tweet possède une URL)
- engagements avec le média (nombre de clics sur le le média)
- ouverture des détails (nombres de vues des détail relatif au tweet)
- clics sur le profil
- clics sur les hashtags
- réponses (nombre de réponse)
- j'aime (nombre de j'aime)

il est possible qu'il existe un nombre d'abonnés rapporté par un tweet mais ce cas ne m'est pas arrivé
]