S4, Développement d’un robot pour extraire des données à partir de Twitter

Nael Boussetta
Quentin DONADONI
Lucas Veynand

I/ Une description détaillée du sujet en mettant l’accent sur les missions attendues par le projet

Nous allons devoir créer un robot qui aura un compte twitter. Celui-ci ira consulter d'autres profils, et grâce à l’outil sélénium, le bot récupérera des informations. Dans un premier temps les tests ne seront faits que sur nos comptes personnels.
Ces informations seront récupérées et stockées dans une base de données. Le robot devra donc créer des bases de données et y stocker les informations récupérées.
Une fois la récolte de données finie sur un profil, le bot doit mettre en fil d’attente tous les followers et les comptes que follow le compte étudié afin d'analyser ces comptes par la suite.
L’objectif final est de pouvoir faire des requêtes sur ces bases de données afin de pouvoir en faire des statistiques.

II/ Une étude de l’existant (recherche web, étude comparative, …)

Il existe une multitude de bot twitter, en effet il y a de nombreux tutoriels qui expliquent comment créer des bots pour des applications diverses. Ceux-ci sont souvent utilisés pour tweeter automatiquement ou pour envoyer des messages automatiques sur mesure.

Parmi les bots twitter existants permettant répondant au mieux à nos problèmes, il y’a :

- **Twexlist** qui sert à extraire et sauvegarder n’importe quelles informations sur Twitter (tweets, abonnés, amis, mentions, membres d’une liste, retweets, favoris, résultats de recherche, messages privés). Il stock les informations dans un fichier Excel.
- **L’API Search Tweets** qui sert à extraire des publications selon une recherche particulière (un ou plusieurs mots-clés, des hashtags ou des noms d’utilisateurs ou encore une période donnée). L’API renvoie alors un JSON (format de données textuelles dérivé du langage JavaScript. Il permet de représenter de l’information structurée comme le permet XML).
- **L’API Get Tweets Timelines** permet notamment de récupérer le fil d’actualités basé sur nos abonnements ou de directement récupérer les 20 derniers tweets d’un utilisateur donné.
- **L’API TwitteR ou R Based Twitter Client** donne accès à l'API Twitter. La plupart des fonctionnalités de l'API sont prises en charge, avec une préférence pour les appels à l'API qui sont plus utiles pour l'analyse des

données que pour l'interaction quotidienne. Cet API est celle officiel de Twitter.

- **Twitter Analytics** fournit toutes les données analytiques de base afin de comprendre comment vos tweets impactent votre communauté. Cela inclut une vue d'ensemble du nombre d'impressions, un classement des meilleurs tweets, l'évolution de votre communauté et le taux d'engagement.

	Twexlist	**Search Tweets**	**Get Tweets Timelines**	**TwitteR**	**Twitter Analytics**
+	Beaucoup d'information extraite	Base de Donnée utilisée Beaucoup d'information extraite	Base de Donnée utilisée	Toutes les fonctionnalités de twitter Beaucoup d'autres fonctionnalité	Est très graphique et donc simple d'utilisation
-	Base de Donnée utilisée		Peu d'information extraite	Pas vraiment d'extraction de données	Est incomplet en version gratuite il faut donc la payer pour plus de détail

Si on les compare, on remarque que TwitterR, l'API officiel de Twitter, est l'API qui se rapproche le plus de notre objectif, cependant celle-ci reste assez limité et ne faisant par exemple par de réel extraction de données.

III/ Une étude technique sur l'existence de solutions permettant de mener à bien le projet

Grâce à l'outil sélénium, nous allons scanner des comptes Twitter sous format HTML afin d'y récolter toutes les données possibles. Puis par la suite, grâce à java nous allons faire du JDBC afin de créer des tables de base de données pour y stocker toute ces données.