Projet M1 Java

M1 Informatique - Algorithmie et Programmation Avancée

Année Universitaire 2019 - 2020

1 Introduction

Twitter est une plateforme de micro-blogging créee en mars 2006 en Californie. Elle permet de publier des messages jusqu'à 280 caractères, de suivre d'autres utilisateurs et d'interagir avec eux à travers deux types d'actions: la mention et le retweet.

Twitter est rapidement devenu un lieu d'échange d'idées, de discussion de sujets sociétaux ou de commentaire de l'actualité. L'analyse des échanges et des communautés qui se créent sur ce réseau sont donc d'un grand intérêt pour beaucoup de chercheurs en sciences sociales.

Néanmoins, l'étude de ces données est extrêmement complexe. 500 millions de tweets sont publiés par jours et 362 millions de personnes utilisent le service par mois [1], nécessitant l'introduction de méthodes d'analyse automatique de données pour explorer, simplifier et organiser l'information.

Dans le cadre de ce projet, vous devrez développer un outil d'exploration de données Twitter. Deux jeux de données tests ont déja été construits et sont disponibles en ligne[2]. Le format des données ainsi que la méthode d'extraction sont détaillés en section 2. Le but est de fournir un outil développé en JAVA pour explorer et analyser ces données. Trois axes d'analyses sont proposés: fouille de texte, fouille de graphe et reporting. Ils seront détaillés en section 3. Les groupes seront libres d'explorer un ou plusieurs de ces axes.

2 Les données

Les données ont été récupérées en utilisant l'API fournie par Twitter[3] pour le requêtage de sa base de donnée. Deux jeux de données ont été créés. Tous les tweets et retweets en français contenants les mots clés "climat", "climatique", "environnement","environnemental" et "environnementaux" sont récupérés depuis le 02/09/2019 , formant un premier jeu de données sur le climat. Similairement, les Tweets contenants les mots clés "foot", "football", "#WWC2019", "#CM2019", "#FootFeminin" et "#FIFAWWC" ont été récupérés du 21/06/2019 au 10/07/2019 pendant la phase finale de la coupe du monde féminine. Les tweets sont stockés au format CSV, séparés par des tabulations ("\t"). Les colonnes sont l'id du tweet, l'id de l'utilisateur, la date de publication, le contenu et l'id de l'utilisateur retweeté si le tweet est un retweet.

3 Objectifs détaillés

Tous d'abord, vous devrez prévoir l'import des jeux de données ainsi que leur stockage. Vous profiterez de la modélisation objets pour faciliter ce stockage. **Peu importe le ou les axes que vous choisissez, vous devrez développer une interface permettant l'import des données**

[1] `https://blog.hootsuite.com/fr/27-statistiques-twitter-a-connaitre-en-2019` `https://blog.twitter.com/marketing/en_us.html`

[2] `https://bul.univ-lyon2.fr/index.php/s/hTA4E2z6h8UtQ9a` mdp:java2019. Attention, la diffusion de ces données n'est pas autorisée sans une anonymisation préalable et aucun usage autre qu'universitaire ne devra en être fait

[3] `https://developer.twitter.com/en/docs.html`

UserID	Date	Texte	RTID
luffy_mh	2019-08-28	La diplomatie climatique d'@EmmanuelMacron échoue à nouveau	greenpeacefr
ecolopress	2019-08-28	Un bilan environnemental mince pour le G7	

Table 1: Exemples de tweets dans leur format de stockage CSV (la date a été tronquée dans les exemples)

3.1 Axe 1: Fouille de texte

Les tweets contiennent énormément d'informations pertinentes pour les sociologues ou les linguistes. Néanmoins, la masse de données rend compliquée leur exploration. Pour cela, vous créerez un outil permettant le requêtage de la base de tweet. Par exemple, un utilisateur devrait pouvoir récupérer tous les tweets contenants les mots "Java", "Université". Ensuite, vous devrez proposer plusieurs fonctionnalités pour simplifier la navigation dans le contenu textuel retourné:

- tri par années
- tri par utilisateurs
- tri par nombre de retweet
- amélioration du requêtage en utilisant des méthodes de normalisation (tf idf par exemple)
- Clustering

3.2 Axe 2: Fouille de graphe

Un graphe est un objet mathématique défini par un ensemble de noeuds, ici les utilisateurs, et un ensemble de liens (arcs dans le cas d'un graphe dirigé, arête sinon) entre ces noeuds. Dans notre cas, les noeud sont les utilisateurs de Twitter, et la relation définissant le lien est le retweet. Concrètement, on créera un lien entre l'utilisateur A et l'utilisateur B si A a retweeté B.

Après avoir construit ce graphe à partir des données fournies, vous pourrez présenter (du plus simple au plus avancé)

- Des statistiques sur ce graphe: volume, ordre, diamètre, degré moyen.
- Les utilisateurs les plus centraux (en terme de Page Rank ou autres mesures de centralité c.f. `https://en.wikipedia.org/wiki/Centrality`.)
- Une visualisation du graphe (vous pourrez filtrer les noeuds)
- Une simplification du graphe en communauté

3.3 Axe 3: Reporting

Ici, la visualisation est centrale et c'est l'analyse temporelle qui prédominera. Le but est de proposer une navigation simple et intuitive dans les jeux de données. Vous proposerez des outils de reporting, en présentant des statistiques simples par périodes (par jours, par semaine, par mois). La visualisation doit être dynamique et modifiable par l'utilisateur.

- Nombre de tweets
- Utilisateurs les plus populaire
- fréquences de hashtags