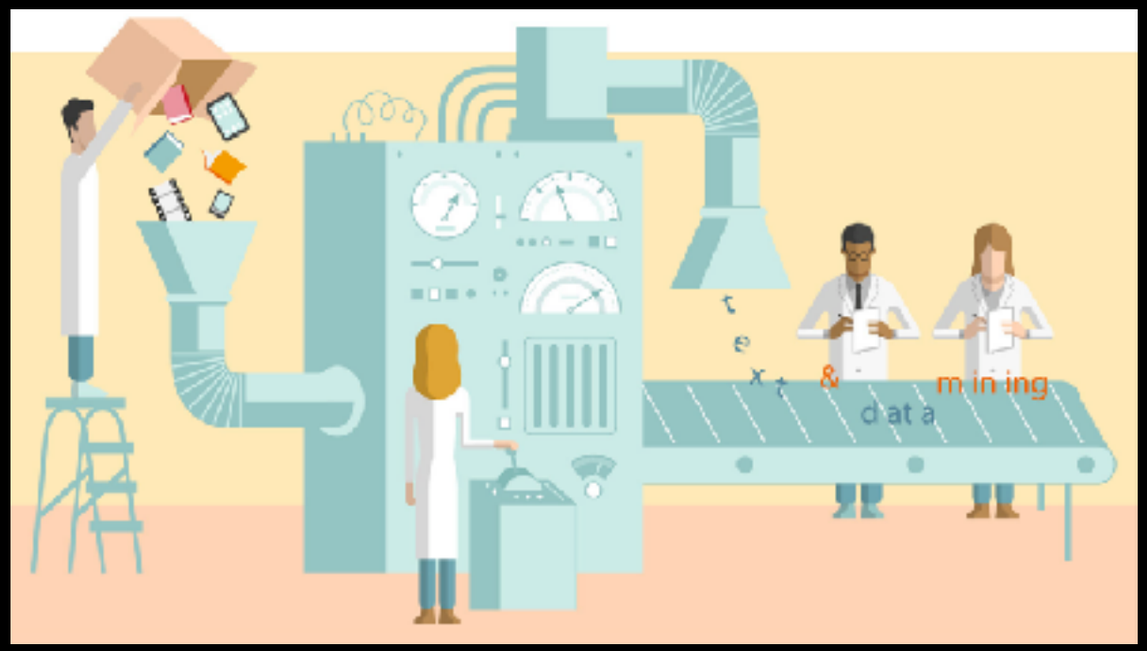


Data Mining & Social Media Robots

Par Raphael Gaudreault

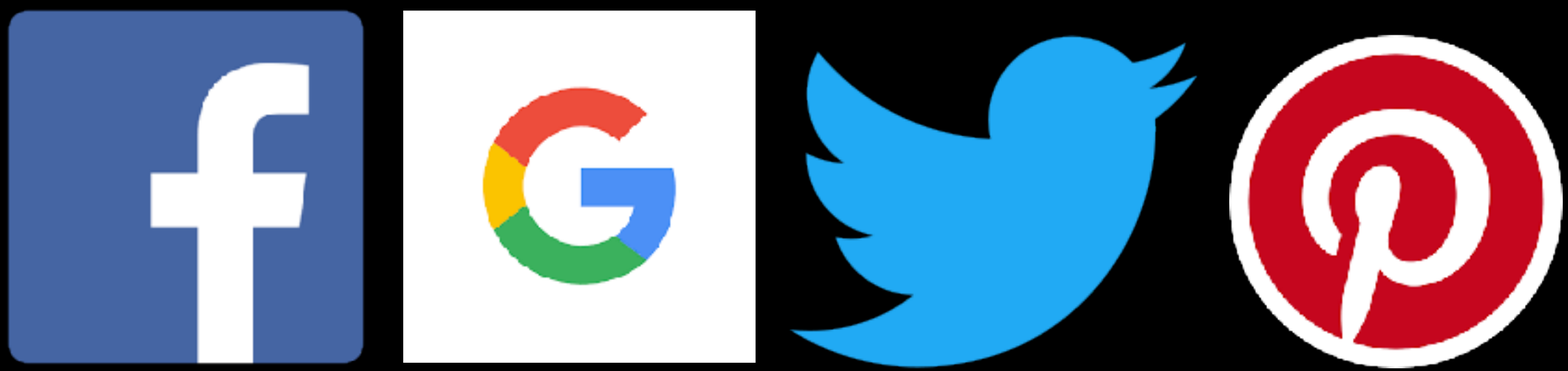
Cette présentation est en franglais

Le data mining en 2017

- “The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use”

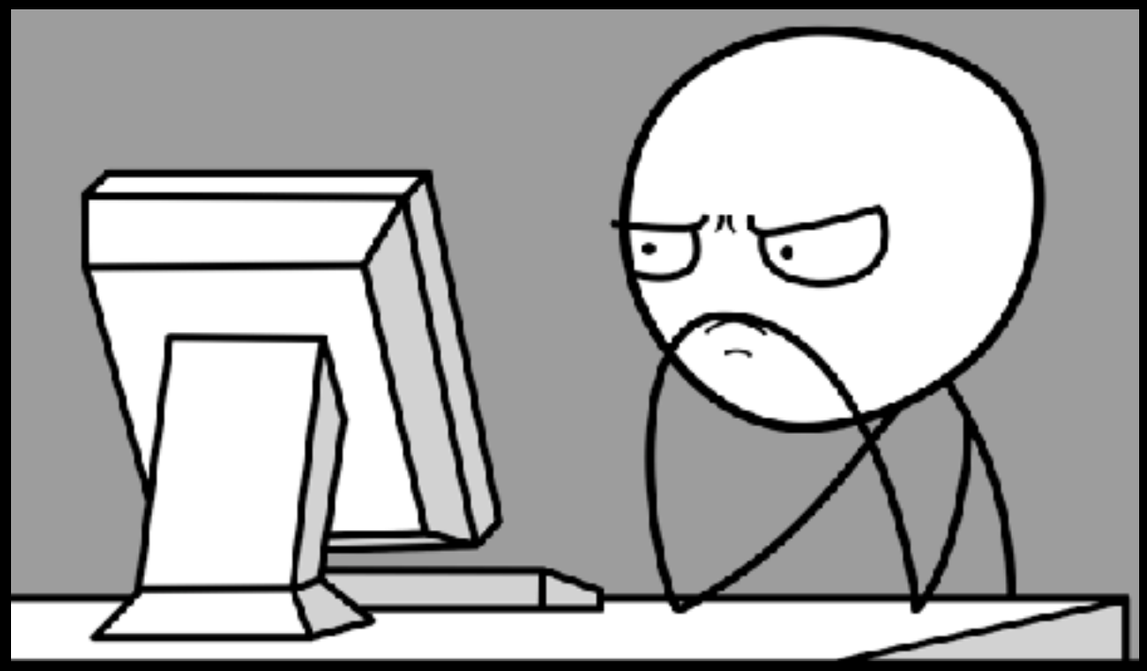
- Est le carburant de l’intelligence artificielle, et je ne vous apprend rien en vous disant ici que c’est l’avenir.

- Le “carburant” de la révolution industrielle 4.0



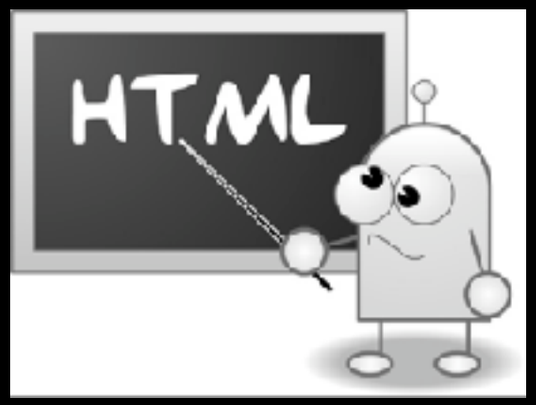
“Vous n’êtes pas le client, mais bien le produit”

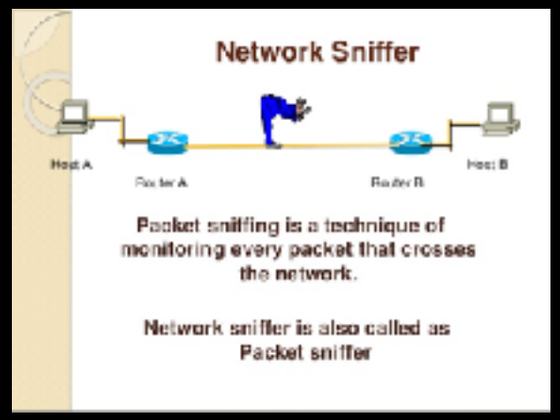
Et vous dans tout ça ?

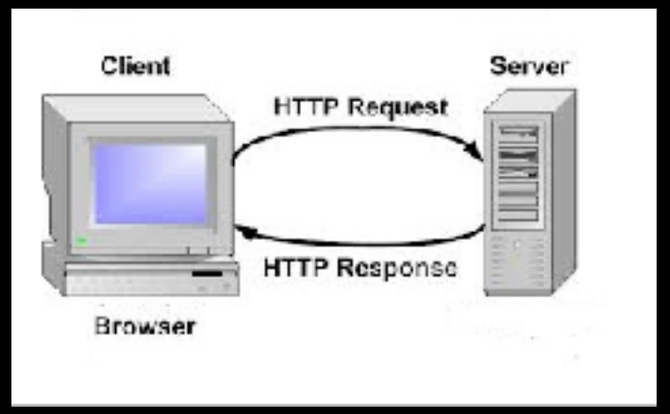


Création d’algorithme pour extraire des données sur la source d’information la plus riche de l’humanité : le WWW

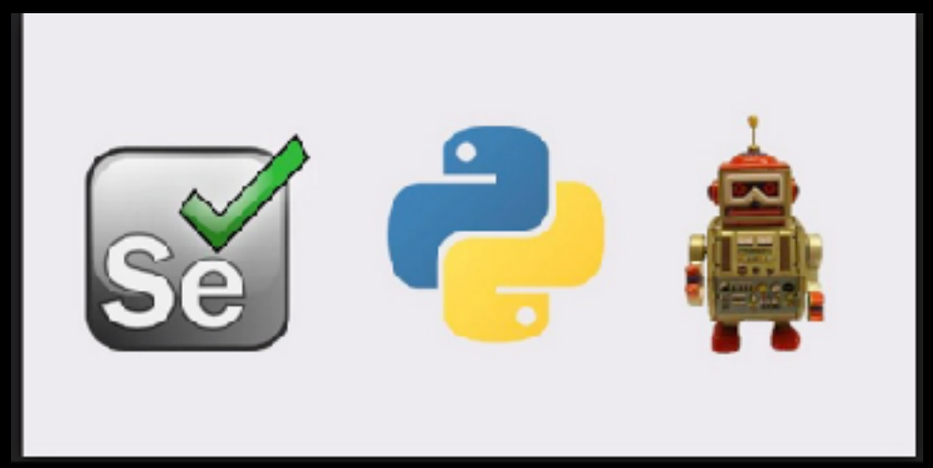
Skills recommandés





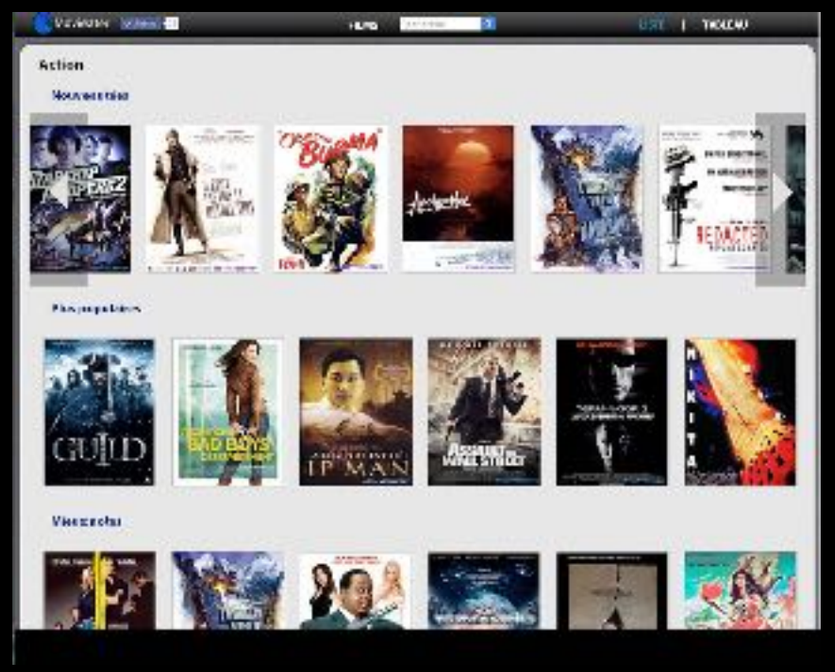


C’est assez simple en fait ! Cela devient plus complexe lorsque le site miné possède des mécanismes protègeant l’intégrité des données.



Selenium -> Vraiment lent

Mon expérience perso



Exemple : Canada411

1. Sniffer trafic http ou https
2. Filtrer les requêtes nécessaires
3. Analyser les calls http (cookies, session etc.)
4. Codage de l’algorithme
5. Faire s’que tu veux avec les données?

Exemple plus complexe : Bot Instagram

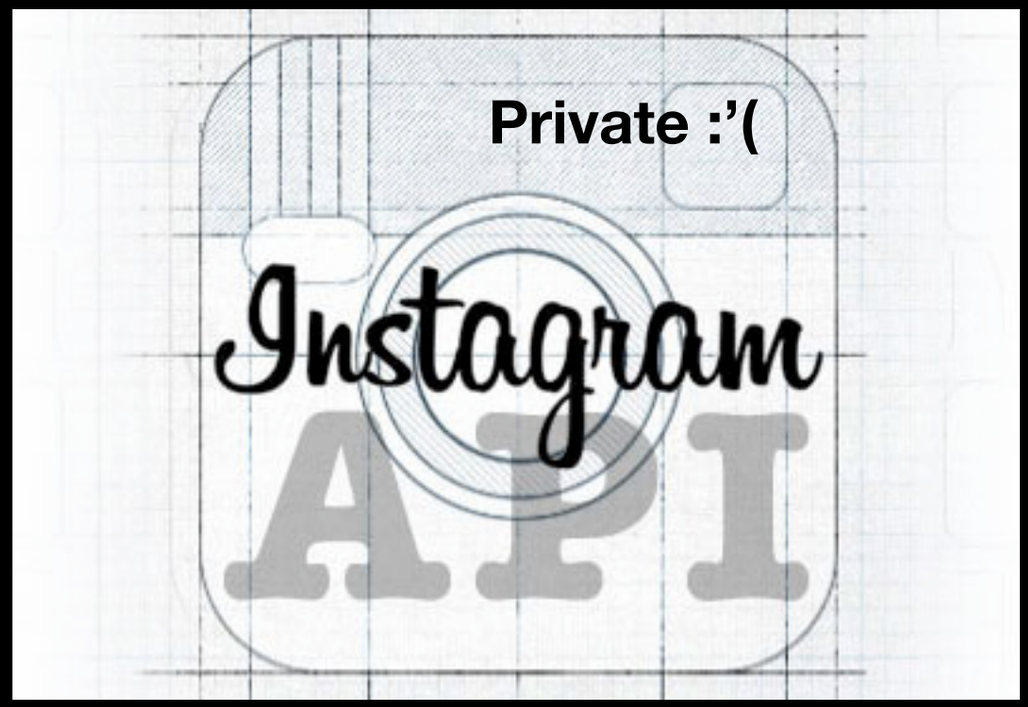
1. Sniffer le traffic https grâce au reverse engineering de l’apk Instagram pour contrer le certificate pinning

HTTP Public Key Pinning (HPKP) is a security mechanism delivered via an HTTP header which allows HTTPS websites to resist impersonation by attackers using mis-issued or otherwise fraudulent certificates. In order to do so, it delivers a set of public keys to the client (browser), which should be the only ones trusted for connections to this domain

Comment ?

- Décompilation de l’APK (.dex) en code java (using jadx)
- Analyse du code client pour contrer le certificate pinning
- Compiler le code modifié en nouvel APK
- Installer cette nouvelle app sur ton cell Android
- Configurer le proxy de ton cell vers ta machine et sniffer le trafic

2. Analyser les requêtes et imiter le behaviour du bot



3. Coder le tout

- Essayer d’être “human-like” le plus possible
- Objet InstagramAPI, Response, etc.
- Traitement des erreurs et escalation
- Architecture Master Server -> Slave servers
- Application Web pour gérer
- Caching des activités de chaque robot

Architecture utilisée

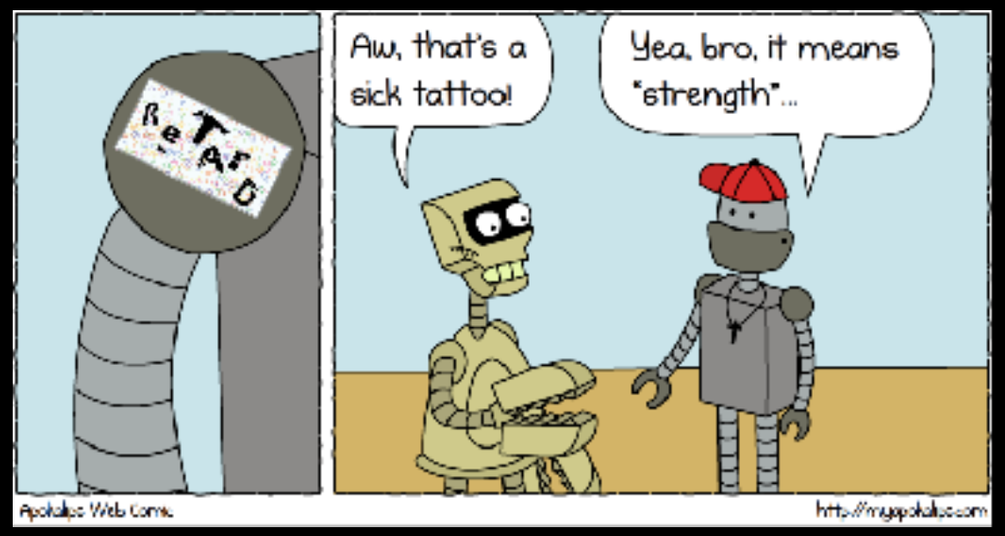
PostgreSQL

redis

python™

Procédures du bot

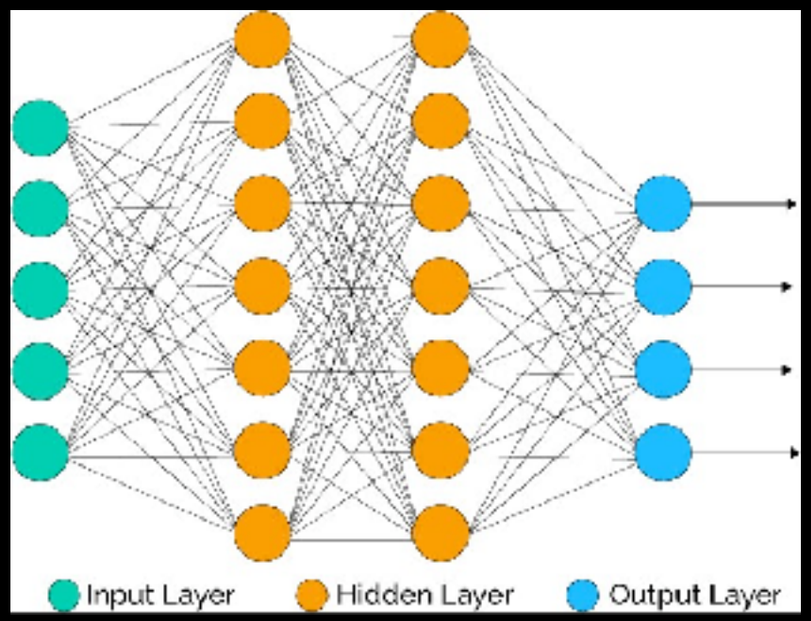
- Post photo from {Reddit, Pinterest, Unsplash}
- Follow specific user
- Register a new account
- Edit profile
- Solve Captcha
- Solve Phone verify



Projet externe utilisé/inspiré

- Reddit Image scraper
- Pinterest board scraper
- Unsplash python API
- Instagram WEB API Python robot
- Instagram private api PHP (old version)

Futur du projet



1.Prendre tous les données de chaque robot et optimiser la fonction de follow back !!!!

2. En amassant les données d’une boutique en ligne, il serait aussi possible d’optimiser la fonction de ventes…

Si il reste du temps: Essayer de sniffer et scraper canada411?

- PS : Je peut assister si vous avez des questions

Merci de votre écoute!