WEB MINING

FARD - Us VS Amazon

Guillaume Hochet - Romain Silvestri - Olivier Kopp

Deux idées
Deux tentatives

Dropshipping

Crawling par recherche inversée d'images sur google

Comparaison des prix de vente pour identifier les sites de dropshipping

Amazon bot detection

Crawling des produits, magasins et avis utilisateurs sur Amazon

Analyse de sentiments sur les avis pour évaluer la qualité de la note donnée

Similarité textuelle pour identifier les bots

Crawler pour dropshipping

Fonctionnement et logique

Requête image ?

Non

Parsing de la page (infos produit)

Recherche devise, comparaison taille police...

Trouver d'autres produits liés

Combinaison de devise, nombre et image à proximité

Récupérer images et liens produits liés

Oui

Ouvrir Google Images

Exécution du code Javascript Google en Headless

Recherche inversée

Collecte des URL des sites

Problèmes

Liés au crawling de multiples sites

- Différences de structure de pages entre les sites
  - Affichage des prix (1'000,32 VS 2,330[30])
  - Image ou background CSS
  - Prix barrés (actions), CSS, SVG...
- Pas d'indexation des sites de dropshipping
  - Google les indexe pas (pas le temps?)
  - Sites créés sur Shopify avec durée de vie courte
- Blocage des requêtes
  - Google Image bloque très rapidement
  - Les différents sites crawlés aussi
- Cohérence des données
  - Conversion de devises (cours pas disponible gratuitement et facilement)
  - Difficulté à être générique

Conclusion sur le dropshipping

Abandonné car pas viable dans le cadre du projet de WEM

Nécessite beaucoup trop de temps

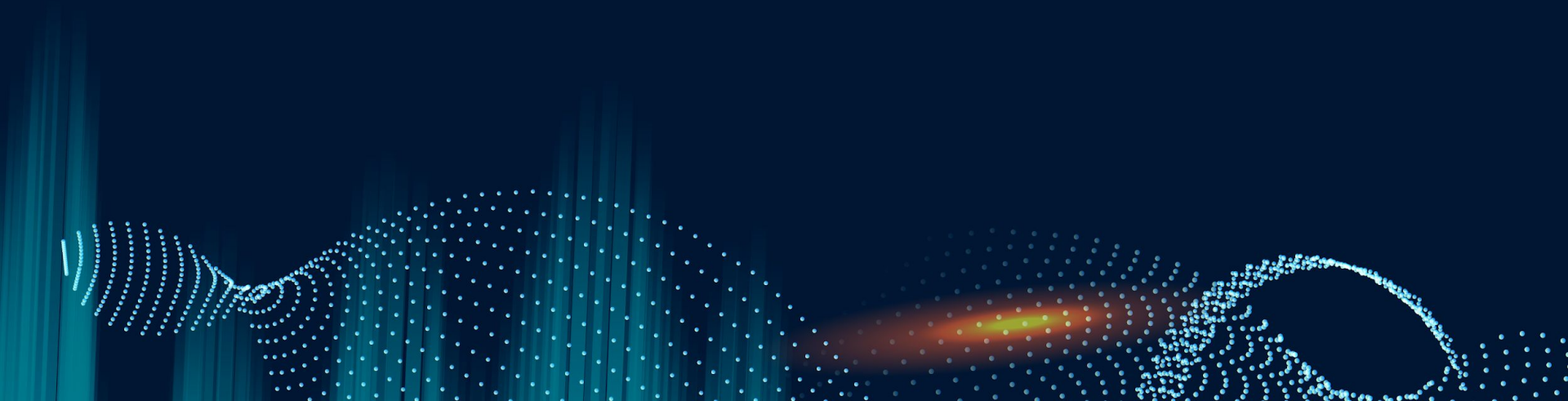
Et de ressources

Quand même quelques résultats intéressants mais pas convaincants

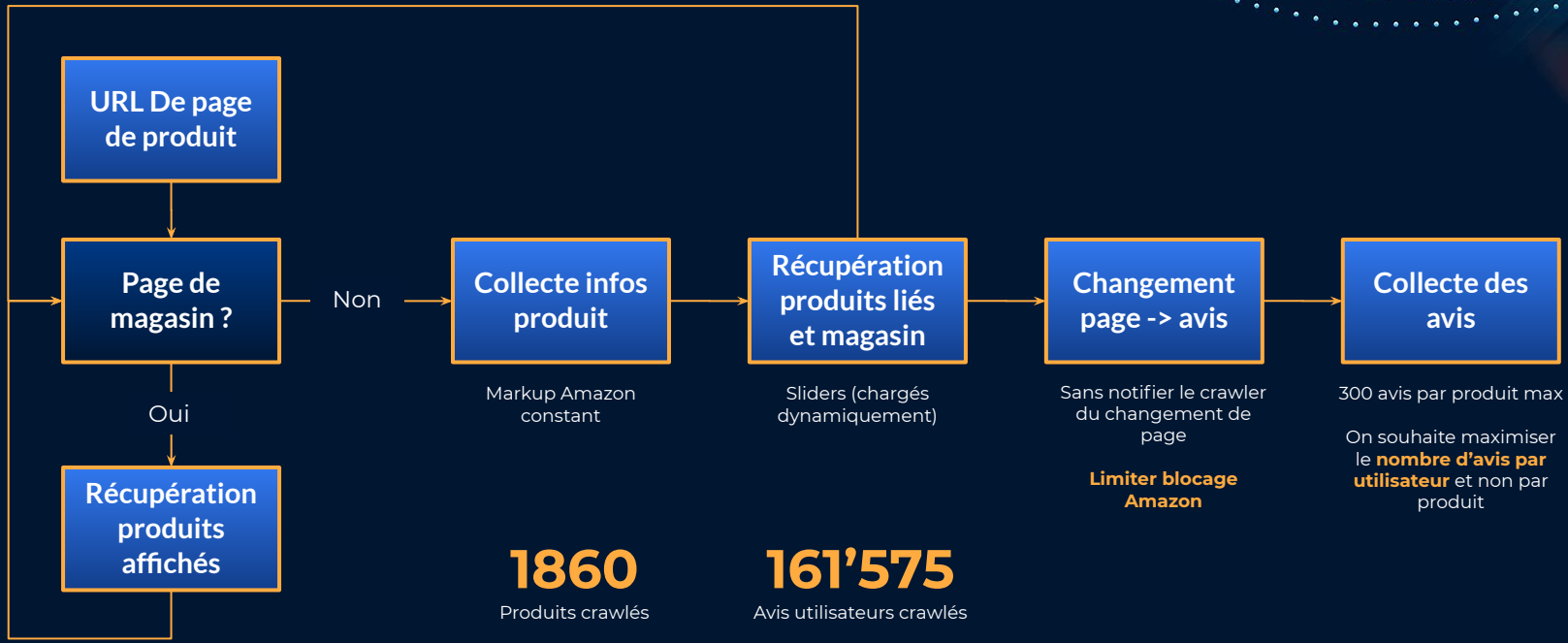
Amazon

Détection de bots et évaluation des avis utilisateurs

- Voir si les notes données par les utilisateurs correspondent à ce qu'ils ont écrit dans leurs avis
- Essayer de repérer les robots qui notent les produits d'une même entreprise



Crawler pour Amazon
Fonctionnement et logique



Blocage par Amazon

Stratégies pour le minimiser

Cliquer sur les boutons pour la navigation

Amazon ajoute des tokens uniques comme paramètres d'URL (type CSRF)

Pas possible de générer l'URL dynamiquement

Scroller les pages dynamiquement

Augmente durée sur une page (crawl plus lent mais plus proche d'un comportement humain)

Permet de charger le contenu dynamique (sliders de produits)

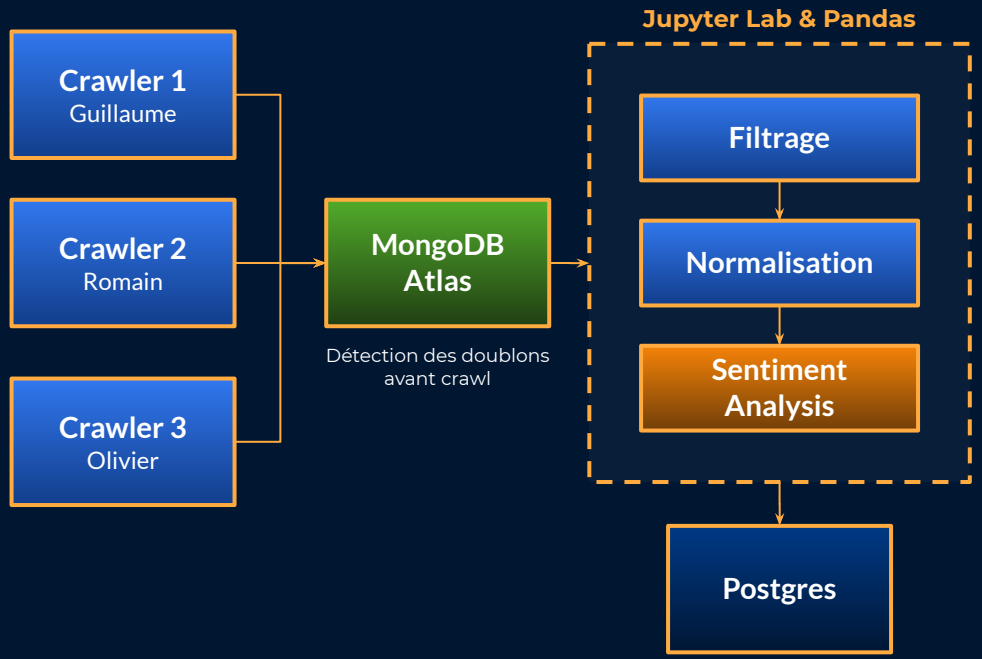
Timeouts aléatoires et bouger le curseur

Timeouts lors des clics sur boutons pour naviguer ou charger du contenu

Plus proche du comportement humain même si très rapide

Pipeline des Données

Collecte et Traitement



Exécutions parallèles du Crawler
minimiser le blocage et maximiser la collecte

Initialement utilisation de RapidMiner
Difficulté à utiliser les données générées

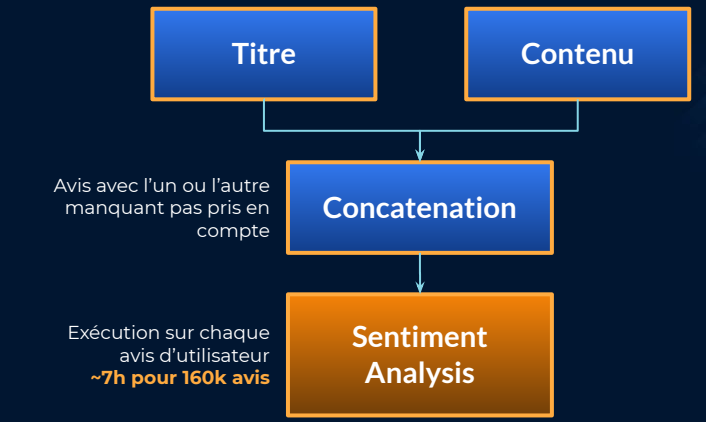
Traitement et processing avec Jupyter Lab & Pandas
Scripts Python

Insertion post-processing dans Postgres
Pour l'application web et l'analyse de données

Sentiment Analysis

Modèle BERT multilingue

- Librairie Python Hugging Face Transformers
- Modèle déjà pré-entraîné
  N'avions pas de données labellisées pour faire du fine-tuning
- Mis à disposition par NLPTown
- Output échelle 1 à 5 (comme Amazon)
- Taux d'exactitude variable
  - 67% correctement prédit
  - 95% correctement avec 1 étoile de décalage



CeraVe Stars Distribution

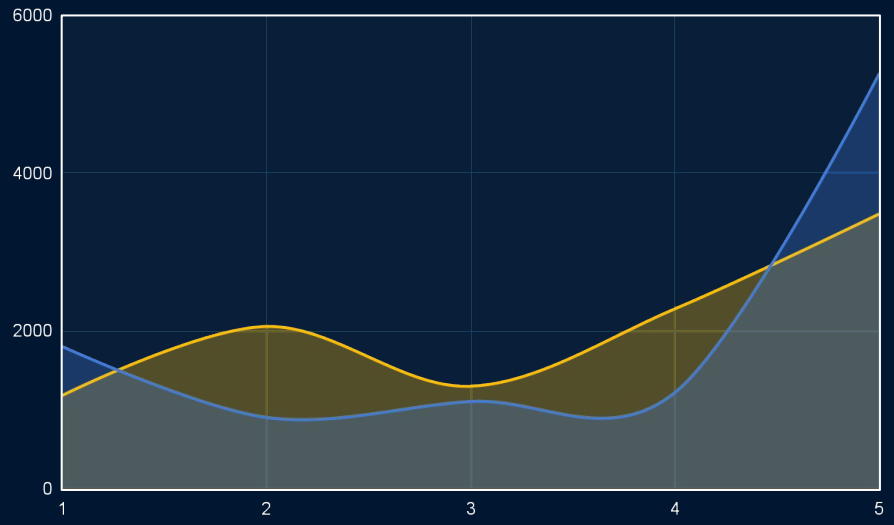
Tendance par entreprise

3.8
Moyenne des avis utilisateurs enregistrés

3.5
Moyenne des résultats d'analyse de sentiments

5148
Avis utilisateurs enregistrés

22
Produits crawlés et enregistrés



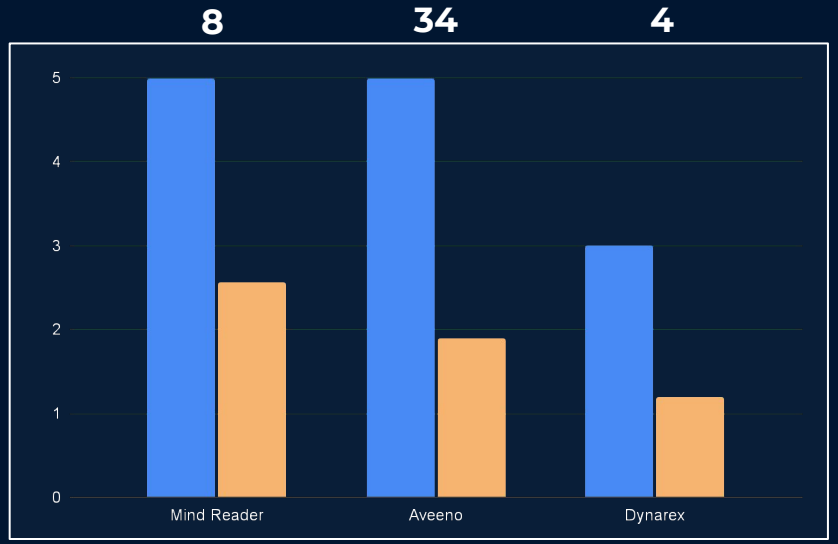
Illustre l'observation générale que nous n'avons pas observé de différences notables dans l'ensemble

Cas particulier

Résultats intéressants sur les produits

34

Avis pour le produit d'Aveeno



Ces 3 produits ont la plus large différence de moyennes entre les avis utilisateurs et l'analyse de sentiment

Détection de bots

String Similarity

Idée - trouver des similarités entre les commentaires rédigés par un même utilisateur

Grouper les avis par similarité et voir si un utilisateur rédige toujours les mêmes commentaires

Première approche - distance de Levenshtein
Pas du tout concluant

Seconde approche - F1-Score
Très moyennement concluant

String Similarity

F1-Score - Exemple avec deux utilisateurs

48	**Oceania** Avis collectés

9 commentaires similaires en **4** groupes

Utilisateur dont nous avons le plus d'avis collectés

Les avis similaires sont carrément identiques

35	**Antigone Walsh** Avis collectés

16 commentaires similaires en **6** groupes

Plus intéressants, les commentaires sont également identiques mais beaucoup sont uniques

Trop peu pour dire si c'est un robot

Très moyennement concluant

Détection de vots

Groupement par entreprise

Idée - Les robots d'une entreprise auront tendance à commenter les produits de cette entreprise

Grouper les avis d'un utilisateur par entreprise dont il a commenté les produits

48	**Oceania** Avis collectés	
18 Magasins commentés		

35	**Antigone Walsh** Avis collectés	
14 Magasins commentés		

Pas concluant

Demo

Petite application web pour explorer les résultats

Conclusion

Très intéressant de pouvoir développer nos crawlers

Plus compliqué que prévu de collecter des données

Des résultats pas concluant majoritairement dû au manque de données

Difficile de se relancer dans un projet différent suite à l'échec du dropshipping

Satisfaisant de développer une application qui illustre nos résultats