11 projets complets et réalistes en Web Scraping & Data Science, adaptés à la formation Master-SDIA

Contexte général :

Dans le cadre de ces projets, l'étudiant devra concevoir une chaîne complète de collecte, traitement et analyse de données réelles issues du **Web**, en utilisant des techniques avancées de **Web Scraping**, de **Data Science** et éventuellement de **Machine Learning** (**ML**).

Chaque projet combine **collecte de données réelles**, **traitement**, **analyse avancée** et parfois **modélisation (Consultez les questions d'orientation à la dernière page de ce document)**.

Objectifs pédagogiques

À l'issue de chaque projet, l'étudiant devra être capable de :

- Analyser la structure d'un site web (HTML, JS, API)
- Implémenter un scraper robuste (Requests / BeautifulSoup / Selenium / Playwright)
- Gérer les contraintes (pagination, AJAX, anti-bot)
- Construire un dataset exploitable
- Réaliser une analyse exploratoire (EDA)
- Appliquer des méthodes de Data Science / ML / NLP
- Présenter les résultats de manière professionnelle

Structure recommandée pour chaque projet :

- Problématique métier
- Étude du site & contraintes légales
- Web Scraping (statique + dynamique)
- Nettoyage et structuration des données
- Analyse exploratoire (EDA)
- Modélisation / NLP / ML
- Visualisation
- Rapport scientifique
- Soutenance

Technologies autorisées :

- Python
- Requests, BeautifulSoup
- Selenium ou Playwright (obligatoire)
- Pandas, NumPy
- Matplotlib, Seaborn
- Scikit-learn / NLP (selon projet)

Contraintes et bonnes pratiques :

- Respect des **conditions d'utilisation des sites**
- Gestion des erreurs et exceptions
- Code commenté et structuré
- Séparation claire des étapes :

  - ✓ **Scraping**
  - ✓ **Nettoyage**
  - ✓ **Analyse**
  - ✓ **Modélisation**

Livrables attendus :

1. **Code source complet** (.py ou .ipynb)
2. **Dataset final (CSV / Parquet)**
3. **Rapport scientifique (PDF)** comprenant :

   - ✓ Introduction & problématique
   - ✓ Méthodologie
   - ✓ Analyse des données
   - ✓ Résultats et interprétation
   - ✓ Limites et perspectives

4. **Présentation orale (10–15 min)**

La répartition des **33** étudiants de la promotion **2025/2026** du master **SDIA** sur les **11** projets disponibles s'effectuera par trinômes. Ainsi, chaque projet sera pris en charge par un groupe de trois étudiants.

Grille d'évaluation (sur 20) :

Critère	Description	Points
Analyse du site web	Étude HTML/JS, choix des outils	1
Web Scraping	Robustesse, gestion du dynamique	4
Qualité du dataset	Nettoyage, structuration	2
Analyse exploratoire (EDA)	Statistiques, visualisations	2
Data Science / ML	Modélisation pertinente	4
Interprétation des résultats	Esprit critique	3
Rapport & présentation	Clarté, rigueur	4
Total		**20**

Toute partie non commentée ou non analysée ne sera pas comptabilisée.

1. Analyse du marché de l'emploi (Indeed / LinkedIn Jobs)

Web Scraping ou Sources :

- Offres d'emploi (poste, entreprise, ville, salaire, compétences)
- Gestion du contenu dynamique (Selenium / Playwright)

Data Science :

- Analyse des compétences les plus demandées
- Clustering des offres par domaine
- Prédiction du salaire (régression)

Livrables :

- Dataset nettoyé
- Dashboard (Matplotlib / Seaborn / Power BI)

2. Analyse des prix et concurrence e-commerce (Amazon / Jumia)

Web Scraping ou Sources :

- Prix, avis, notes, vendeurs
- Pagination et anti-bot

Data Science :

- Évolution des prix dans le temps
- Analyse des avis (NLP – sentiment analysis)
- Détection de produits sur/sous-évalués

Bonus :

- Système de recommandation simple

3. Prédiction du succès d'un produit e-commerce

Web Scraping ou Sources :

- Avis clients, notes, prix, catégorie

Data Science :

- Feature engineering
- Classification (produit populaire / non populaire)
- NLP sur commentaires clients

Outils :

- BeautifulSoup + Selenium
- Scikit-learn / NLP

4. Analyse des tendances immobilières

Web Scraping ou Sources :

- Sites immobiliers (prix, surface, localisation)

Objectifs :

- Prédire le prix d'un bien
- Comparaison des villes/quartiers
- Cartographie géographique (GeoPandas)

5. Scraping et analyse de données sportives

Web Scraping ou Sources :

- Résultats, statistiques joueurs/équipes

Data Science :

- Analyse de performance
- Classement dynamique
- Modèle de prédiction de résultats

6. Analyse des avis clients sur les hôtels / restaurants

Web Scraping ou Sources :

- Booking / TripAdvisor

Data Science :

- NLP : sentiment, mots-clés
- Score de satisfaction global
- Comparaison par ville/catégorie

7. Veille technologique automatique

Web Scraping ou Sources :

- Blogs, sites d'actualité tech, GitHub

Objectifs :

- Extraction des articles
- Topic Modeling (LDA)
- Détection des tendances émergentes

8. Analyse des réseaux sociaux (X / Reddit / Forums)

Web Scraping ou Sources :

- Posts, commentaires, likes

Data Science :

- Analyse de sentiment
- Détection de sujets populaires
- Analyse temporelle de buzz

Respect des conditions d'utilisation

9. Système de recommandation de formations en ligne

Web Scraping ou Sources :

- Coursera / Udemy (titres, notes, prix, durée)

Objectifs :

- Clustering des formations
- Recommandation personnalisée
- Visualisation des parcours d'apprentissage

10. Surveillance automatique de prix et alertes

Web Scraping ou Sources :

- Suivi quotidien des prix

Data Science :

- Détection d'anomalies
- Prévision des baisses de prix
- Système d'alertes (email / dashboard)

11. Optimisation et Prédiction des Prix Aériens par Web Scraping et Modélisation Avancée :

Web Scraping & Data Science :

- **Analyse Concurrentielle** : Scraper et comparer systématiquement les prix et les offres de plusieurs compagnies sur une route donnée pour identifier des stratégies de pricing.
- **Modélisation Avancée** : Prédire non seulement le prix, mais aussi la **probabilité qu'un prix baisse ou augmente** dans les jours à venir.
- **Recommandation Personnalisée** : Développer un prototype d'application qui conseille le "meilleur moment pour acheter" sur un trajet donné, basé sur l'analyse historique et en temps réel.

Questions d'orientation pour les étudiants :

Phase 1 : Analyse du site

- Quelle est la structure HTML principale des pages ciblées ?
- Les données sont-elles chargées statiquement ou dynamiquement ?
- Existe-t-il une API cachée ?
- Quels sont les risques de blocage (anti-bot) ?

Phase 2 : Web Scraping

- Pourquoi avoir choisi Requests / Selenium / Playwright ?
- Comment gérez-vous la pagination ?
- Comment gérez-vous les erreurs HTTP ?
- Quelle est la fréquence optimale de scraping ?

Phase 3 : Nettoyage des données

- Quelles colonnes contiennent des valeurs manquantes ?
- Comment traitez-vous les doublons ?
- Quelles transformations ont été nécessaires ?
- Pourquoi ces choix de nettoyage ?

Phase 4 : Analyse exploratoire (EDA)

- Quelles sont les statistiques descriptives principales ?
- Quelles variables sont les plus influentes ?
- Existe-t-il des corrélations significatives ?
- Quelles visualisations sont les plus pertinentes ?

Phase 5 : Data Science / ML

- Quel problème avez-vous modélisé (classification, régression…) ?
- Quelles features ont été sélectionnées et pourquoi ?
- Quel modèle donne les meilleurs résultats ?
- Comment interprétez-vous les performances obtenues ?

Phase 6 : Analyse critique

- Quelles sont les limites de votre dataset ?
- Quels biais peuvent affecter les résultats ?
- Comment améliorer ce projet à grande échelle ?
- Quelles applications réelles sont possibles ?

Phase 7 : Approfondissement (bonus)

- Peut-on automatiser la mise à jour des données ?
- Comment sécuriser un scraper en production ?
- Peut-on déployer un dashboard interactif ?
- Peut-on intégrer ce projet dans un SI réel ?
- Quelles considérations éthiques faut-il prendre en compte ?