Université Paris 1 Panthéon-Sorbonne

UFR02 Économie

Master 2 MoSEF

RÉSOLUTION AUTOMATIQUE DE CAPTCHAS VISUELS

Webscraping & API with Python

Techniques : Webscraping, Computer Vision, VLM

Déploiement : API FastAPI

Livrables : Notebook + Rapport + GitHub

Année universitaire 2025-2026

Objectif du projet

Le but du projet est de concevoir un webscraping robuste aux **CAPTCHAs visuels** provenant de différents sites web. Il devra être fait en Python en suivant les bonnes pratiques de code et devra être fait en collaboration sur GitHub entre les membres du projet.

Travail attendu

- construire un **webscraping** robuste aux **CAPTCHAs** ;
- entraîner ou utiliser un modèle permettant de reconnaître le texte ou le contenu du CAPTCHA ;
- concevoir une **API** pour orchestrer et industrialiser le projet. Cette API devra avoir une architecture pertinente, claire et propre et suivre les bonnes pratiques vues en cours.

Livrables

- **Notebook** présentant :
  - la démarche suivie ;
  - les essais réalisés ;
  - les visualisations utiles ;
  - l'évaluation du système.
- **Rapport Overleaf** comprenant :
  1. Méthodologie générale ;
  2. Description du jeu de données ;
  3. Résultats obtenus et interprétation.
- **Dépôt GitHub**

  Attention : nous prendrons en compte l'historique des pushes sur GitHub et la contribution de chaque membre de l'équipe au projet. Un seul push avant le rendu par une seule personne n'est pas acceptable. Avoir des PRs et des reviews sera très apprécié.

Datasets open-Source recommandés

Voici une liste de jeux de données open-source que vous pouvez utiliser pour entraîner et tester votre système de résolution automatique de CAPTCHAs.

Nom du Dataset	Description
LCSD Captcha Dataset	Environ 6 000 images de CAPTCHA (4 caractères) provenant d'un service public, avec labels. Diversité moyenne, utile pour un premier modèle OCR.
Captcha Object Detection Dataset	Environ 100 000 images et 650 000 objets annotés (lettres et chiffres). Convient pour les modèles de segmentation/détection et pour apprendre la structure d'un CAPTCHA complexe.
Pixel Digit Captcha Data	Petits CAPTCHAs composés de chiffres uniquement (5 digits). Dataset simple et efficace pour débuter avec la reconnaissance de texte.
CAPTCHA Characters Dataset	Plus de 118 000 images individuelles de caractères (lettres + chiffres). Utile pour l'entraînement d'un modèle de reconnaissance caractère par caractère.
CAPTCHA Image Dataset	10 001 images CAPTCHA complètes et annotées. Format standard, très pratique pour tester différents modèles OCR.
OpenCaptchaWorld Dataset	Collection variée de CAPTCHAs textuels et non-textuels. Adapté pour développer un système robuste capable de généraliser à plusieurs styles visuels.

TABLE 1 – Jeux de données open-source utilisables pour l'entraînement et l'évaluation d'un résolveur de CAPTCHAs.