EXTRACTION DES DONNEES A PARTIR DES IMAGES

AVOSSE Yaovi Ametepé & GUINARD Océane

Université Paris Cité

May 19, 2023

Plan

Introduction

Définition et problématique

Il n'existe pas un logiciel libre permettant d'extraire automatiquement des tableaux depuis des images.

Introduction

Objectifs

1. Collecte et génération automatique d'un ensemble d'images de tableaux de différents formats;
2. Installation et utilisation d'un algorithme d'OCR existant sur une grande variété d'images représentante des tableaux dactylographiques;

Introduction

Objectifs

1. Collecte et génération automatique d'un ensemble d'images de tableaux de différents formats;
2. Installation et utilisation d'un algorithme d'OCR existant sur une grande variété d'images représentante des tableaux dactylographiques;
3. Enrichir un algorithme d'OCR existant pour extraire des tableaux de format csv;
4. Étudier les performances des modèles neuronaux d'un OCR en spécifiant leur apprentissage sur des ensembles d'images représentant des tableaux.

Introduction

Notre solution

Ce que notre logiciel fait:

- Notre solution ne marche pour l'instant que sur les images simples.
- notre solution consiste à extraire des données sous formats .csv à partir d'images de tableaux.

Architecture

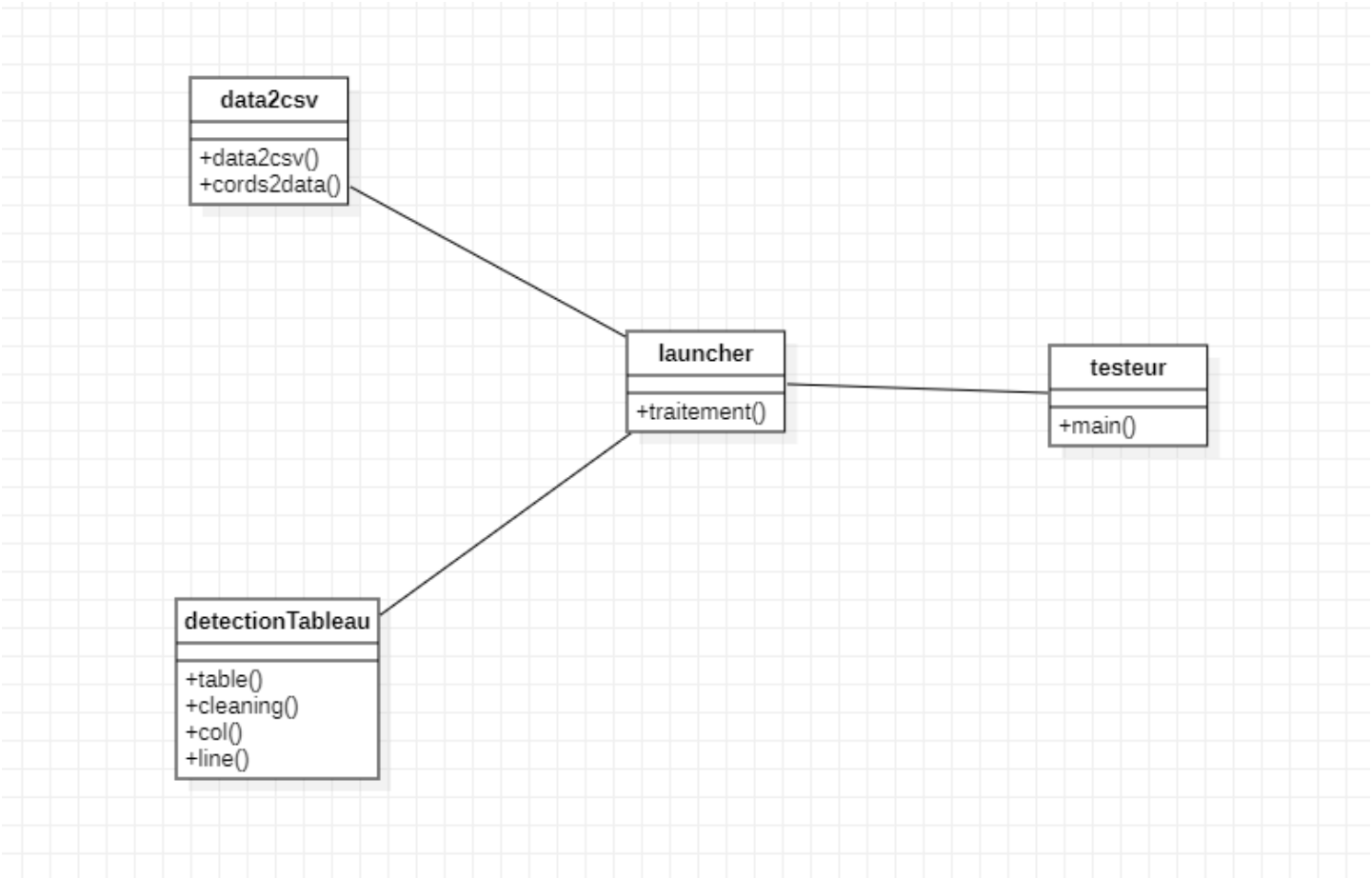

Figure 1: Diagramme d'architecture

Architecture, conception et gestion de projet

Les étapes de la réalisation:

1. openCv récupère chaque case de notre tableau et renvoie ses coordonnées;
2. Nous nettoyons ainsi les lignes et les colones;
3. Après, nous renvoyons ces nouvelles données à Tesseract qui grâce à python reconstitue les cases une par une et transforme le résultat en .csv

Architecture, conception et gestion de projet

Les compétences techniques:

- Python
- pytesseract
- opencv
- numpy, etc.

Les difficultés rencontrées

1. Tesseract a du mal avec les petites cases, mais nous avons pu lui donner un argument qui lui dit que c'est qu'une ligne de texte et depuis cela a amélioré drastiquement ses perfomances;
2. Tesseract détecte des caractères spéciaux en fin de ligne sans raison, nous avons éviter ceci en nettoyant ces caractères spéciaux à chaque fois;

Les difficultés rencontrées

1. Tesseract a du mal avec les petites cases, mais nous avons pu lui donner un argument qui lui dit que c'est qu'une ligne de texte et depuis cela a amélioré drastiquement ses perfomances;
2. Tesseract détecte des caractères spéciaux en fin de ligne sans raison, nous avons éviter ceci en nettoyant ces caractères spéciaux à chaque fois;
3. Tesseract reconnaît plus ou moins bien certaines polices d'ecriture;
4. openCv a beaucoup de mal à detecter les cases si celles-ci ne sont pas bien definis;
5. Dans le cas de nos tableaux, Tesseract renvoie une information qui la plupart du temps necessiterait une petite modification pour obtenir le tableau attendu.

Programmation

- Détection des cases du tableau via OpenCV.

Programmation

Nettoyage des cases inutiles:

- premier algorithme trival, on supprime les cases vide;
- probleme:

	matière	note
	PLONG	
	sys	13

Figure 2: Tableau avec case vide

Programmation

- Deuxième algorithme: vérifier que les bords du tableau soit vide.

Testabilité

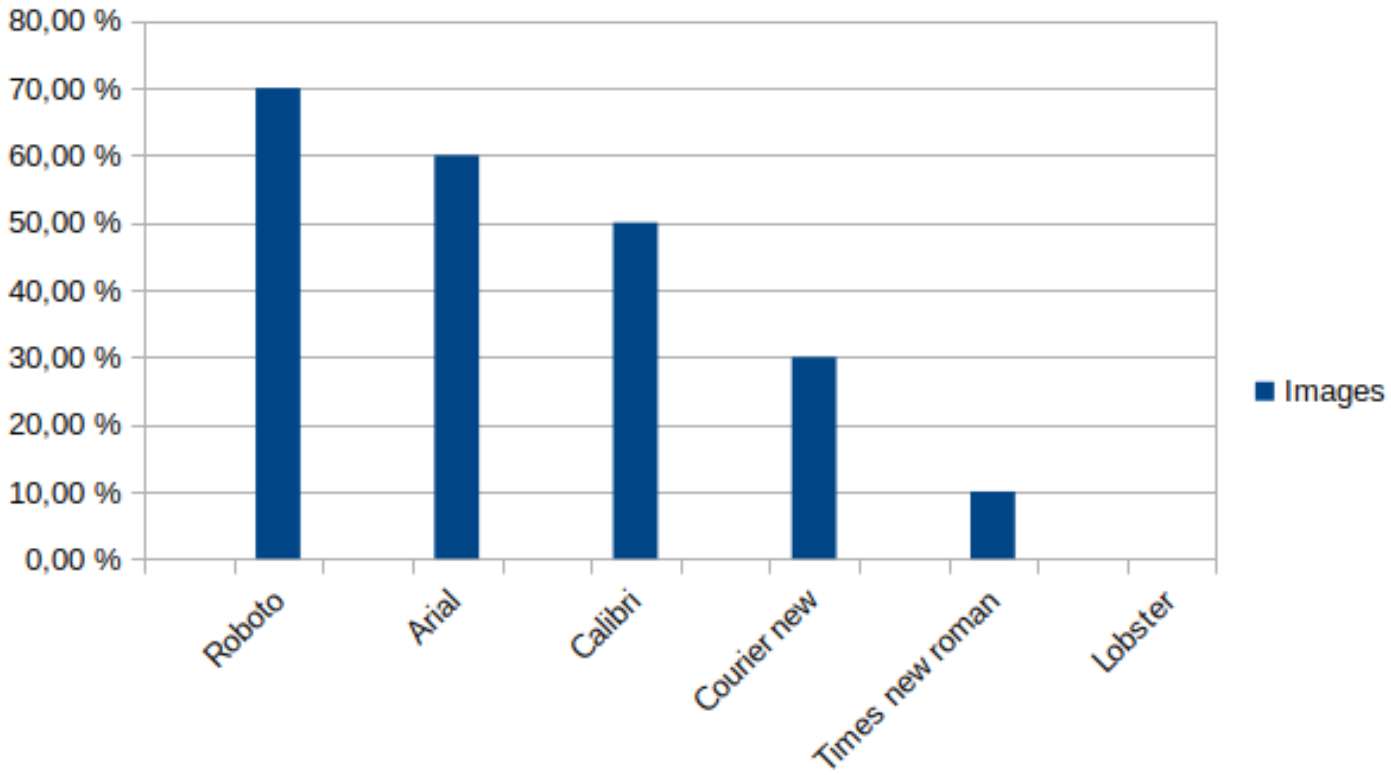

Figure 3: Performance de notre solution

Conclusion

Nous avons beaucoup appris de ce projet:

1. comment combiner tesseract et opencv pour faire une bonne extraction des données;
2. Une version 2.0 de ce projet prendra en compte les tableaux compacts et les images contenant des tableaux en couleurs.

Conclusion

- Si on devait refaire le projet maintenant, on aurait chercher à entraîner un algorithme d'OCR sur une police bien précise.

Merci pour votre attention!