REVUE DE PROJET

Équipe 2

[redacted] Marion

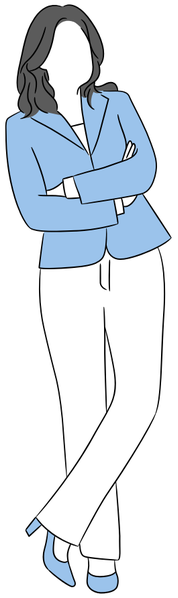
Master Industrie de la Langue – 2024/2025 – Gestion de Projet

SOMMAIRE

1

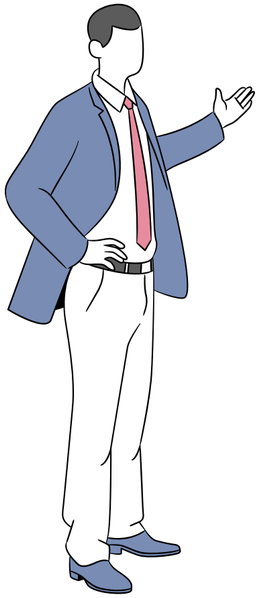
1

Présentation de l’équipe



[redacted]

- **Développement informatique**



[redacted]

- **Développement pédagogique**

Marion [redacted]

- **Recherches documentaires et constitution de corpus**

1

2

Organisation et outils de travail

Communication & échange de documents

Backlog sur Excel, mis à jour régulièrement

Développement des scripts

2

1

L’origine du projet : un manque

Constat :

Peu (voire pas) de cours centrés sur l’OCR dans notre formation de TAL

Alors que :

L'OCR est une technologie essentielle dans de nombreux domaines : numérisation de documents, analyse de textes manuscrits, reconnaissance de caractères, etc.

Notre solution :

Créer une formation pour combler ce manque de connaissances sur un sujet aussi important, pour donner aux étudiants en TAL une base plus solide pour aborder le marché du travail

+ Proposer une introduction à l’OCR à un public plus large que les étudiants en TAL

2
2

Notre solution : combler ce manque

La matérialisation de notre solution :

Un cours en ligne sur la plateforme open-source Moodle, disponible pour les étudiants en TAL (mais pas seulement)

moodle

Introduction aux systèmes d'Optical Character Recognition

La reconnaissance optique de caractères (OCR) est une technologie essentielle permettant la conversion d'images ou de documents scannés en texte exploitable. Ce cours en ligne propose une introduction à l'OCR, de son histoire et son évolution à ses applications modernes dans divers domaines (archivage, automatisation, accessibilité, etc.)

À travers des explications théoriques et des exercices pratiques, vous apprendrez à :

- Comprendre le fonctionnement de l'OCR et son importance.
- Explorer les différentes technologies et solutions disponibles (avec l'exemple concret du modèle *EasyOCR*)
- Installer et utiliser un modèle OCR sur votre propre machine.
- Entraîner un modèle OCR sur des jeux de données spécifiques pour l'adapter à vos besoins.
- Évaluer les performances des systèmes OCR et améliorer leur précision.

Ce cours s'adresse aux débutants souhaitant découvrir l'OCR ainsi qu'aux développeurs et data scientists qui veulent approfondir leurs connaissances sur l'entraînement et l'évaluation des modèles OCR.

Pré-requis

- Connaissances de base en programmation (Python recommandé)
- Notions en manipulation d'images et en intelligence artificielle (un plus mais pas obligatoire)

2

3

Formulation du projet selon les critères SMART

Publier avant le 17 avril, un cours en ligne sur Moodle pour permettre aux étudiants en TAL de combler leurs lacunes sur l'OCR. Le cours, structuré en modules pratiques et théoriques, leur apprendra à installer, utiliser, entraîner et évaluer un modèle OCR localement, efficace à au moins 70% selon le *Character Error Rate*.

3 Les parties prenantes du projet

Partie prenante	Interne/ Externe	Influence	Intérêt	Réactivité
Plateforme	Externe	Forte	Faible	Faible
Utilisateur	Externe	Faible	Fort	Faible
Équipe projet	Interne	Forte	Fort	Moyen

4 Les personae

Plateforme pensée pour les étudiants en TAL, et éventuellement d'autres mais avec des bases en informatique

→ Personae en conséquence

Cours pensé <u>pour des étudiants</u> : utilisateurs en formation continue =/= pas le public visé

Clara Dupont

Lucas Martin

- **Clara Dupont, 23 ans, Master en TAL**
- **Solide formation en linguistique théorique et appliquée**
- **S'intéresse particulièrement aux applications de l’IA dans le traitement des données textuelles et des langues naturelles**

<u>Objectifs d'apprentissage :</u>

- **Souhaite comprendre comment l’OCR peut être utilisée dans des projets linguistiques, notamment pour traiter des documents manuscrits**
- **Veut se familiariser avec les outils et bibliothèques qui permettent de réaliser des systèmes OCR performants afin de pouvoir les intégrer dans ses projets de recherche**

- **<u>Style</u> : Autodidacte, apprécie une approche progressive avec des modules théoriques, suivis d’applications pratiques**
- **<u>Préférences</u> : Préfère les vidéos courtes et les articles scientifiques détaillant les méthodes d'implémentation + aime aussi les exercices pratiques, où elle peut tester et appliquer directement les concepts**
- **<u>Motivations</u> : Motivée par des projets de recherche qui lui permettent de combiner ses intérêts pour les langues et les technologies avancées + souhaite accroître ses compétences pour son futur travail en laboratoire de recherche ou dans l'industrie.**

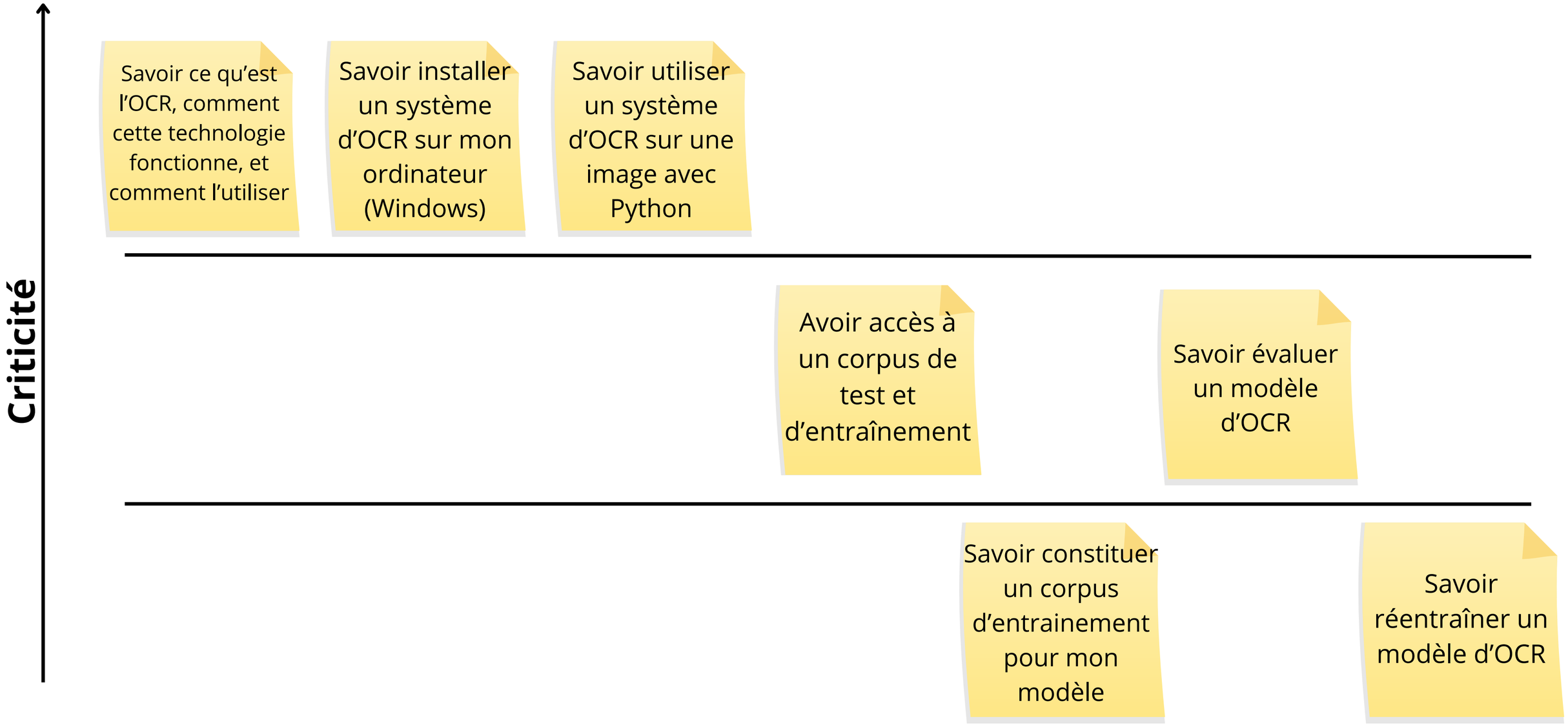
- **Lucas Martin, 26 ans, Licence d’Histoire, autodidacte en IA**
- **Parcours académique en sciences humaines**
- **Passionné d’IA et d’OCR et leur application pour la numérisation de documents historiques**

<u>Objectifs d'apprentissage :</u>

- **Comprendre les bases de l’OCR et des modèles d’apprentissage automatique sans avoir de formation initiale en TAL ou en IA**
- **Apprendre à utiliser des outils et bibliothèques OCR existants (Tesseract, EasyOCR, PaddleOCR) pour ses projets personnels**
- **Savoir entraîner et ajuster un modèle OCR pour reconnaître des textes anciens et des manuscrits difficiles à numériser**

- **<u>Style</u> : *Trial & error*, en suivant des tutoriels et en expérimentant avec du code + il préfère des explications claires et des démonstrations interactives**
- **<u>Préférences</u> : Aime les vidéos explicatives et les blogs techniques accessibles aux débutants, avec des cas concrets et des lignes de code faciles à adapter**
- **<u>Motivations</u> : Veut pouvoir restaurer et analyser des documents anciens numérisés et rêve de développer un projet personnel dans ce domaine**

5 La retranscription des besoins en objectifs : le story-mapping



5 Les besoins utilisateur et les contraintes

Priorités Client

- **Pouvoir lancer un système d'OCR peu importe mon sys. d'exploit.**

Services Client:

- **Avoir des scripts de ré entrainement pour tous les sys.exp.**

Fonctions User

- **Savoir constituer un bon corpus d'OCR**
- **Savoir lancer un système d'OCR sur une image avec Python**

Services User

- Avoir accès à un corpus et à des modèles pré-entraînés
- Avoir accès à la formation en ligne

Fonctions Prod

- **Être en capacité de publier une formation**
- **Connaître les exigences des clients pour publier une formation**

Conditions Prod

- **Modèles OCR open-source.**
- **Corpus de 10 000 images.**

Contraintes Prod

- **Temps limité: 2 mois.**
- **Ressources techniques nécessaires.**

5 Le backlog

Rédaction du backlog en fonction des besoins de chaque partie prenante

Exemple :
En tant que : U*tilisateur*
J'ai besoin de : *Connaître les critères théoriques de constitution d'un corpus pour un modèle d'OCR*
Afin de : *Pouvoir évaluer la pertinence du corpus choisi*

+ Expression de *Critères d'acceptation* pour que chaque membre de l'équipe ait une idée claire de quoi faire

En tant que	J'ai besoin de	Afin de	Critère d'acceptation
Utilisateur	Savoir comment évaluer mon modèle d'OCR comparativement à d'autres modèles d'OCR pour une même tâche	Pouvoir choisir le modèle le plus performant en sachant justifier mon choix	- Avoir trois modèles prêts à l'emploi validant les exigences du besoin n°2 et avoir un des modèles ayant une performance de 70 % de f-mesure sur la partition de test du corpus

5 Le backlog

Indication temporelle / en personnel pour chaque livrable, + échéance

Estimation	Statut	Nom du livrable	Echéance
1/2 personne/2 semaines	Tâche validée	Test de trois modèles sous python	26/02/25

6 Bilan RSE

Impact environnemental	EasyOCR : open source, plus léger et moins énergivore	Moodle : serveur local (ordinateur personnel) → pas d'hébergement sur serveur qui consommerait beaucoup d'énergie	Utilisation de ChatGPT pour les personae et certains morceaux de code → empreinte carbone non-négligeable
Impact social	EasyOCR : très inclusif (80+ langues reconnues dans plusieurs alphabets différents)	Cours proposé uniquement en français, et pour Windows → public limité en termes de langue et de connaissances	Peu de réflexion sur le TTS, descriptions alternatives des images systématiquement proposées
Éthique et gouvernance	Données libres de droit	Sources claires et partagées dans le cours, pas de plagiat	Guide d'utilisation éthique de l'OCR ?
Propositions d'actions RSE	Certification "éthique" en fin de formation	Proposer plus d'options d'accessibilité à la formation	Limiter autant que possible l'usage de l'IA générative

6 Bilan RH

Répartition des rôles	**→ Selon les appétences de chacun** **Marion et Sonia** : recherche documentaire **Lou** : aspect technique **Antonin** : conception de la formation
Compétences développées	**TAL :** Constitution de corpus, évaluation modèles OCR, utilisation d'outils (EasyOCR) **Transversales :** Travail en équipe, organisation sur 2 mois, adaptabilité
Gestion de la charge de travail	Répartition entre tâches **techniques** (prétraitement, scripts) et tâches **linguistiques / pédagogiques** (constitution corpus, création tutoriel)

6 Bilan RH

Communication et collaboration	WhatsApp pour le suivi, Google Docs pour partage de documents → **écoute active**
Motivation et bien-être	Contraintes de temps avec d'autres cours et partiels → réorganisation des priorités
Forces et faiblesses	**+** : Complémentarité des profils (SDL ou autre) et autonomie **–** : Communication parfois imprécise