TAD : Traitement Automatique de Documents : APPLICATION à l'OCR ARABE

Projets à réaliser

Réglementation

1. Les projets peuvent réalisés au minimum en binômes et au maximum de six étudiants.
2. Les projets peuvent être exposés devant l'ensemble des étudiants.
3. Le choix du projet est laissé libre aux étudiants.
4. Les projets avec les mêmes contenus se verront attribuer un 'zéro'.

Dates

La date **limite** de choix des projets est prévue pour le

Vendredi 30/09/2022.

Dates

La date **limite** de remise des projets est prévue pour le

Vendredi 18/11/2022.

A remettre

1. Un rapport format Word détaillant le travail effectué.

2. Un code source en python.

3. Une interface d’utilisation.

4. Une présentation format PowerPoint.

Intitulés

I. Traitement Automatique de Documents

1. Détecter les zones des images
2. Détecter les zones écrites
3. Segmentation de l’écriture en lignes
4. Segmentation en mots

Intitulés

II. Filtrage, binarisation

1. Réaliser une interface qui permet :
   a. De choisir une suite de filtres pour un bon rendu de l'écriture.
   b. De permettre de paramétrer le filtrage par l'utilisateur.
2. Proposer une technique adaptative interactive de binarisation de l'écriture.

Intitulés

III. Linéarisation dans le domaine cartésien

1. Segmentation de l'écriture arabe en parties connexes.
2. Implémentation de l'algorithme de squelettisation « Zhang & Suen ».
3. Proposer une technique de linéarisation de l'écriture : S'inspirer des directions de Freeman pour les directions à adopter.

Intitulés

IV. Linéarisation avec une structure d'arbres n-aires

1. Segmentation de l'écriture arabe en parties connexes.
2. Implémentation de l'algorithme de squelettisation « Zhang & Suen ».
3. Représentation des parties connexes dans des structures d'arbres n-aires.
4. Proposer une technique de linéarisation de l'écriture en utilisant la même structure : S'inspirer des directions de Freeman pour les directions à adopter.

Intitulés

V. Orthogonalisation dans le domaine cartésien

1. Segmentation de l'écriture arabe en parties connexes.
2. Implémentation de l'algorithme de squelettisation « Zhang & Suen ».
3. Proposer une technique d'orthogonalisation de l'écriture : on prend les deux directions horizontale et verticale.

Intitulés

VI. Orthogonalisation

1. Segmentation de l'écriture arabe en parties connexes.
2. Implémentation de l'algorithme de squelettisation « Zhang & Suen ».
3. Représentation des parties connexes dans des structures d'arbres n-aires.
4. Proposer une technique d'orthogonalisation de l'écriture en utilisant la même structure : : on prend les deux directions horizontale et verticale.

Intitulés

VII. Segmentation

1. Proposer une technique de segmentation en parties connexes.
   a. Appliquée à l'écriture contrainte
   b. Appliquée à l'écriture non contrainte
2. Séparer les parties principales des diacritiques et proposer une technique de segmentation en graphèmes avec affectation des diacritiques aux graphèmes appropriés.

Intitulés

VIII. Extraction de caractéristiques

1. Implémenter trois méthodes d'extraction de caractéristiques.
2. Implémenter une méta-heuristique de votre choix pour sélectionner les caractéristiques de chaque méthode, puis concaténer les trois ensembles résultants.
3. Valider la méthode de sélection avec un classifieur de votre choix.

Intitulés

IX. Extraction de caractéristiques

1. Implémenter trois méthodes d'extraction de caractéristiques.
2. Implémenter une méta-heuristique de votre choix pour sélectionner les caractéristiques de l'ensemble des trois méthodes, combinées de façon ordonnée, puis de façon aléatoire.
3. Valider la méthode de sélection avec un classifieur de votre choix.

Intitulés

X. Combinaison de classifieurs

1. Choisir une méthode d'extraction de caractéristiques à implémenter.
2. Choisir trois classifieurs à utiliser.
3. Proposer une technique de combinaison de ces classifieurs.

Intitulés

XI. Deep learning

1. Proposer une architecture de DL pour la reconnaissance du manuscrit arabe traitant une image de mot entier.

2. Étudier l'effet de quelques techniques d'augmentation des données

Intitulés

XII. Utilisation de Tesseract

1. Installation de Tesseract et Pythesseract
2. Configuration pour travailler avec la langue arabe
3. Utilisation de Python-OpenCV-Pytesseract pour reconnaître des images avec écriture arabe : imprimée, manuscrite contrainte, manuscrite non contrainte.

Intitulés

XIII. Utilisation de EasyOCR

1. Installation de EasyOCR
2. Configuration pour travailler avec la langue arabe
3. Utilisation de Python-OpenCV- EasyOCR pour reconnaître des images avec écriture arabe : imprimée, manuscrite contrainte, manuscrite non contrainte.

Intitulés

XIV. Utilisation de Keras-OCR

1. Installation de Keras-OCR
2. Configuration pour travailler avec la langue arabe
3. Utilisation de Python-OpenCV- Keras-OCR pour reconnaître des images avec écriture arabe : imprimée, manuscrite contrainte, manuscrite non contrainte.

Remarques

- Concernant les trois derniers projets, sachant que les outils considérés sont open-source, toute proposition d'amélioration est la bienvenue.