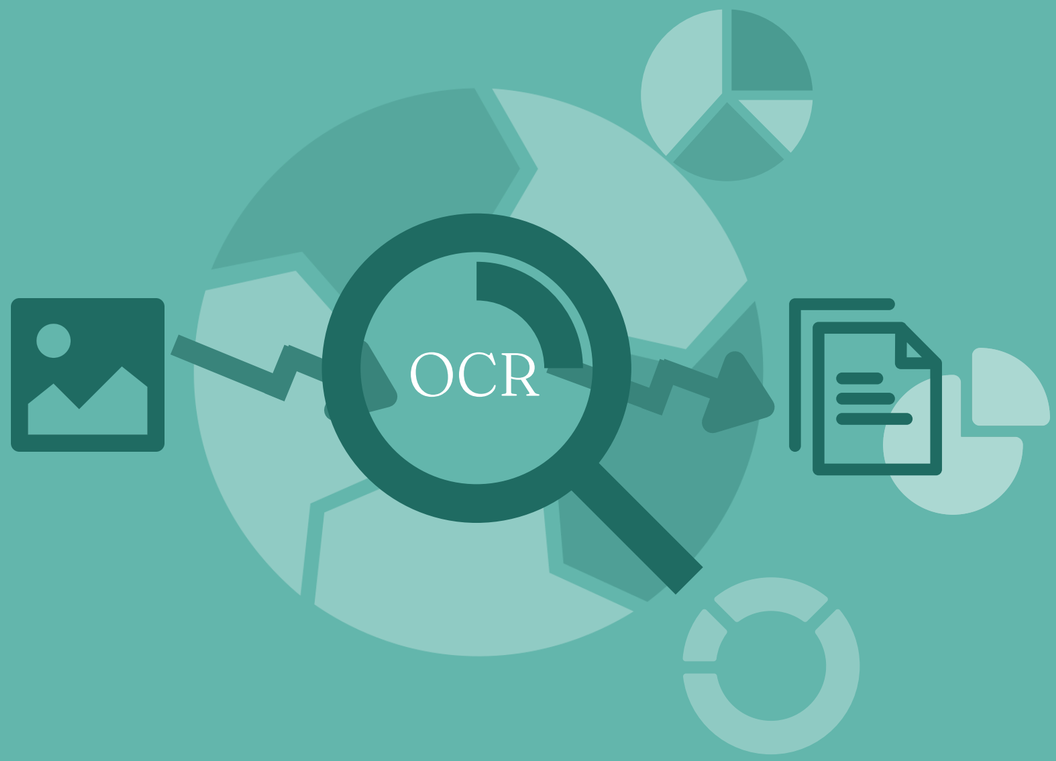
Université de Paris

Extraction d'un tableur depuis une image



Haithem ZEGGARI - Paul Minguet

Encadré par :

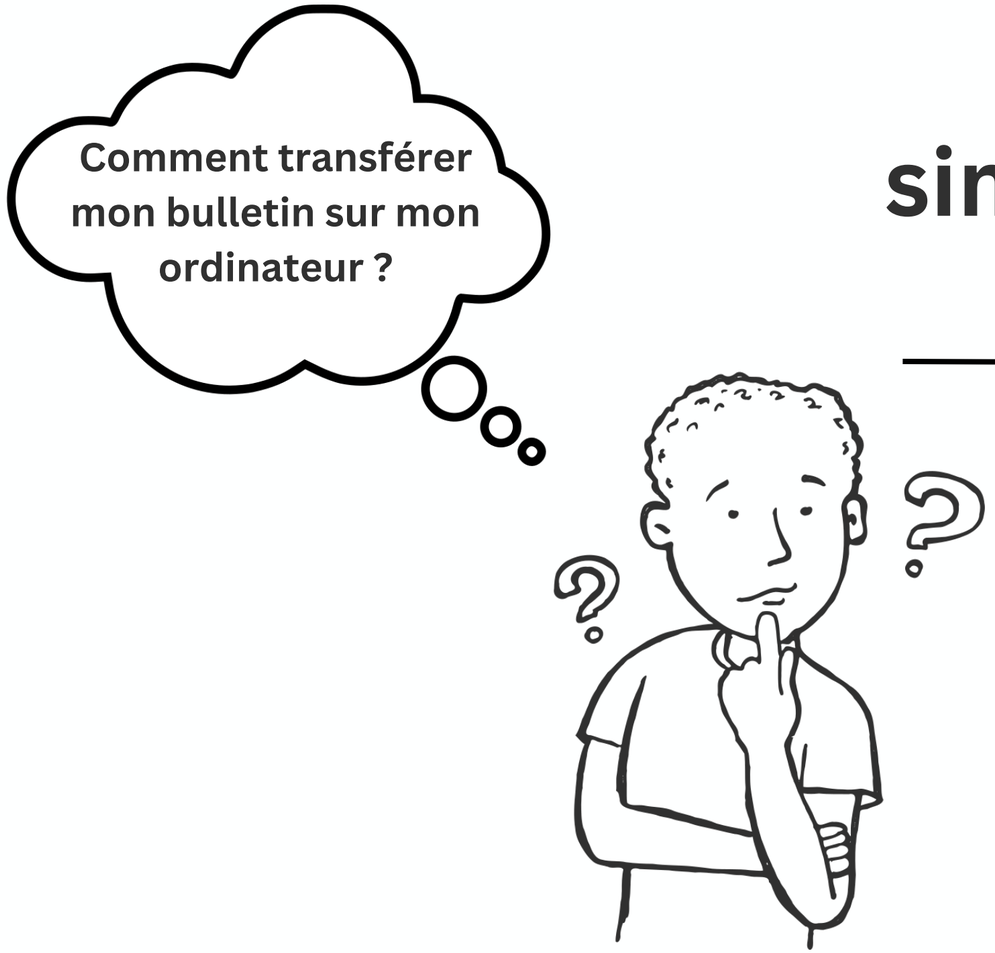
- M. PAGANI
- M. FÉRÉE
- M. DEGORRE

Sommaire

Introduction

Optical Caracter Recognition

Problématique



Comment numériser simplement un tableau sur papier ?

Objectifs

- **Création d'un algorithme pour extraire un tableur depuis une image**
- **Créer un jeu de données conséquent**
- **Étudier les performances de l'algorithme**

Démonstration

Calendrier

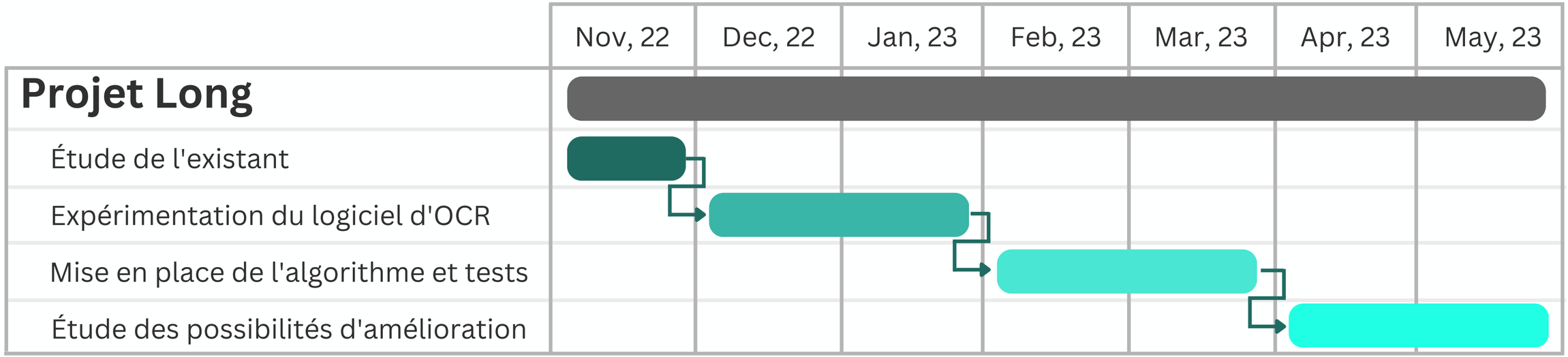
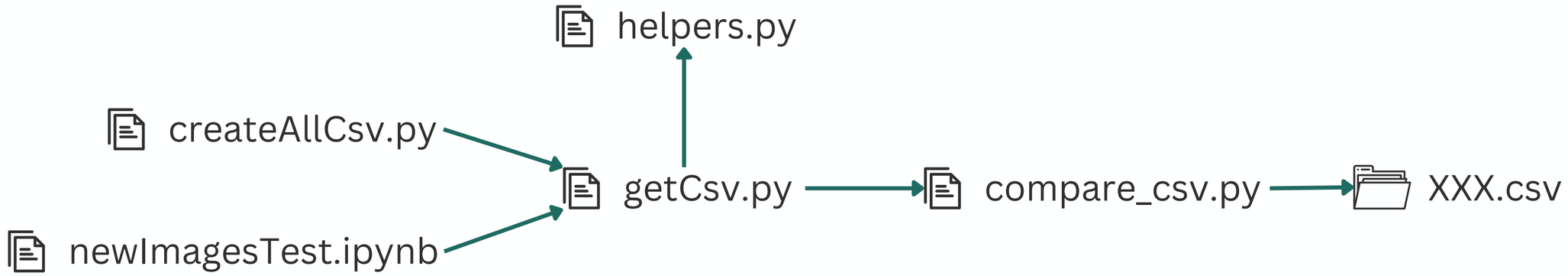


Diagramme de Gantt du projet

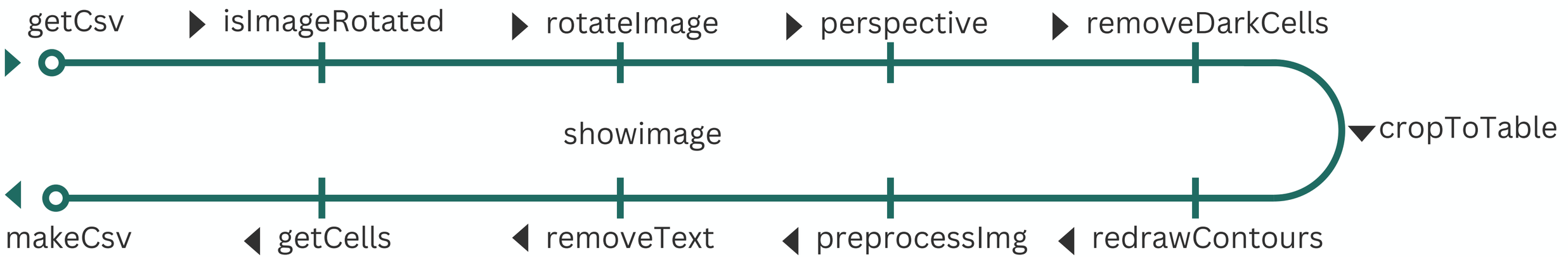
Architecture du projet

- zeggari-minguet-plong-2022
  - .vscode
  - installation
  - tesseract
    - image
      - easy-noColors
        - 01.csv
        - 01.jpg
        - 02.csv
        - 02.jpg
        - ...
      - medium-colors
      - hard-holes
      - harder-colors
      - scanned
    - observings.md
    - outputs
  - compare_csv.py
  - createAllCsv.py
  - getCsv.py
  - helpers.py
  - newImagesTest.ipynb
  - tests.md
  - usage.md
- journal.org
- README.md

Architecture du projet



getCsv.py


getCsv
isImageRotated
rotateImage
perspective
removeDarkCells
cropToTable
showimage
makeCsv
getCells
removeText
preprocessImg
redrawContours

Difficultés

Compréhension du fonctionnement de Tesseract OCR

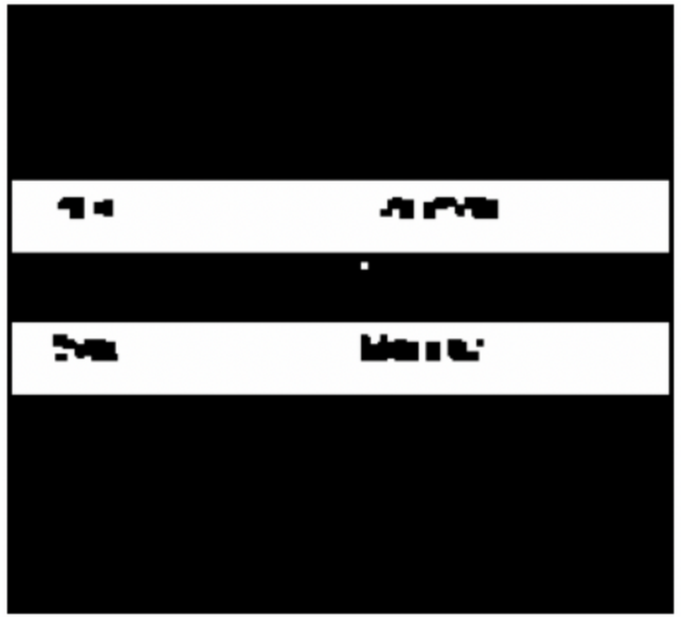
Recherche des fonctionnalités d'OpenCV

Trouver des solutions pour implémenter certaines fonctionnalités

Récupération des coordonnées des cellules

Nom	Prénom
Daniel	Alves
Fabio	Canavarro
Lio	Messi
Riyad	Mahrez
Cézar	Azbelicueta
Santos	Da Silva
Ali	Ahmed

Tableau à tester



Image"morph"

Nom	Prénom
Daniel	Alves
Fabio	Canavarro
Lio	Messi
Riyad	Mahrez
Cézar	Azbelicueta
Santos	Da Silva
Ali	Ahmed

Cases inversées

Récupération des coordonnées des cellules

Nom	Prénom
Daniel	Alves
Fabio	Canavarro
Lio	Messi
Riyad	Mahrez
Cézar	Azbelicueta
Santos	Da Silva
Ali	Ahmed

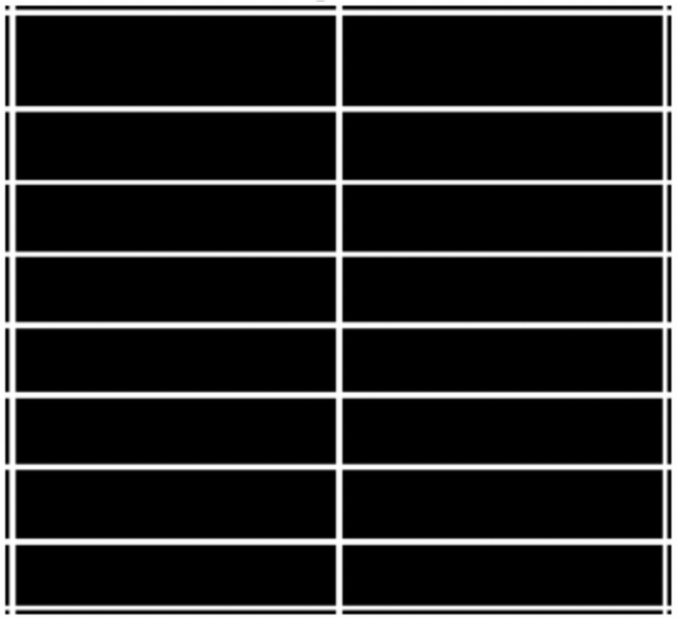
Contours recréés

Nom	Prénom
Daniel	Alves
Fabio	Canavarro
Lio	Messi
Riyad	Mahrez
Cézar	Azbelicueta
Santos	Da Silva
Ali	Ahmed

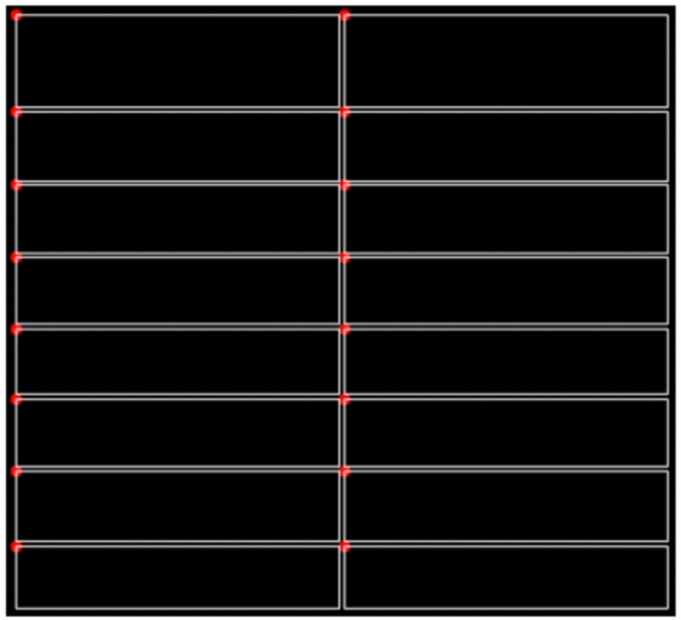
Image pretraitée

Tableau sans le texte

Récupération des coordonnées des cellules



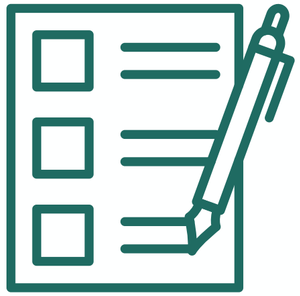
Grille du tableau



Coordonnées des cases

Testabilité

- **Tests sur le jeu de données grâce à un script**
- **Tests à la main sur une image**



Résultats

NOM	PRENOM
Hench	John
Atencio	X
Bob	Gurr
Coats	Claude
Baxter	Tony
Blair	Mary
Rohde	Joe
Sherman	Richard
Sherman	Robert
Davis	Marc
Davis	Alice
Sklar	Marty

Différences :

- "X" -> ""

Pourcentage de similitude :
96.15%

Résultats

Nom	Prénom	Adresse
Daniel	Alves	25 rue Albert eighnshtein 85632
Fabio	Canavarro	35 Rue Bauchaumont, 75002, Paris, France
Lio	Messi	Gare de l'est
Riyad	Mahrez	17 Bd de Strasbourg, 94200
Cézar	Azbelicueta	-
Santos	Da Silva	-
Ali	Ahmed	-

Différences :

- "-" -> ""

Pourcentage de similitude :
85.00%

Résultats

Nom	Prénom	Phone	Email
Daniel	Alves	(258) 879-6320	Dani.alves@gmail.com
Fabio	Canavarro	(102) 787-0023	Fabicanav@gmail.com
Lio	Messi	(222) 589-9623	Lio122-m@gmail.com
Riyad	Mahrez	(213) 751-8794	Mahrezr26@gmail.dz
Cézar	Azbelicueta	(102) 259-6324	C_azb@gmail.com
Santos	Da Silva	(569) 032-7474	santosbr@gmail.com
Ali	Ahmed	(256) 254-0121	aliahled@gmail.com

Pourcentage de similitude :
100%

Résultats

Nom	Prénom
Daniel	Alves
Fabio	Canavarro
Lio	Messi
Riyad	Mahrez
Cézar	Azbelicueta
Santos	Da Silva
Ali	Ahmed

Différences :

- **"Mahrez" -> ""**

Pourcentage de similitude :
93.75%

Résultats

Nom	Prénom
Daniel	Alves
Fabio	Canavarro
Lio	Messi
Riyad	Mahrez
Cézar	Azbelicueta
Santos	Da Silva
Ali	Ahmed

Pourcentage de similitude :
100%

Résultats

Nom	Prénom
Daniel	
	Canavarro
Lio	
	Mahrez
Cézar	
	Da Silva
Ali	Ahmed

Pourcentage de similitude :
100%

Résultats

NOM	PRENOM
Hench	John
	X
Bob	
Coats	
Baxter	Tony
Rohde	Joe
	Richard
	Robert
Davis	
Davis	Alice
Sklar	Marty

Différences :

- "X" -> ""

Pourcentage de similitude :
96.15%

Résultats

NOM	PRENOM	EMAIL
Hench	John	
Atencio	X	
		bgurr@wed.com
Coats	Claude	
Baxter	Tony	tbaxter@wed.com
	Mary	
Rohde	Joe	jrohde@wed.com
Sherman	Richard	dsherman@wed.com
Sherman		rsherman@wed.com
	Marc	mdavis@wed.com
Davis	Alice	
Sklar		lsklar@wed.com

Différences :

- "X" -> ""
- "@wed" -> "@vwed"
- "Tony" -> "on"
- "Marty" -> "Ma rty"

Pourcentage de similitude :
50.00%

Résultats

NOM	PRENOM	PHONE
Hench		
Atencio	X	(839) 589-6673
		(567) 426-7919
Coats	Claude	(708) 343-5400
Baxter	Tony	(783) 719-1744
Blair		(666) 449-1166
	Joe	
Sherman		(410) 431-2226
Sherman	Robert	
Davis		(630) 886-1584
	Alice	(654) 348-9009
Sklar	Marty	

Différences :

- "X" -> ""
- "Tony" -> "on"
- "Marty" -> "Ma rty"

Pourcentage de similitude :
92.30%

Résultats

Nom	Prénom
Daniel	Alves
Fabio	Canavarro
Lio	Messi
Riyad	Mahrez
Cézar	Azbelicueta
Santos	Da Silva
Ali	Ahmed

Différences :

- **Le texte clair n'est pas pris en compte**

Pourcentage de similitude :
56.25%

Résultats

Nom	Prénom	Phone	Email
Daniel	Alves	(258) 879-6320	Dani.alves@gmail.com
Fabio	Canavarro	(102) 787-0023	Fabicanav@gmail.com
Lio	Messi	(222) 589-9623	Lio122-m@gmail.com
Riyad	Mahrez	(213) 751-8794	Mahrezr26@gmail.dz
Cézar	Azbelicueta	(102) 259-6324	C_azb@gmail.com
Santos	Da Silva	(569) 032-7474	santosbr@gmail.com
Ali	Ahmed	(256) 254-0121	aliahled@gmail.com

Différences :

- **Le texte clair n'est pas pris en compte**

Pourcentage de similitude :
2.50%

Résultats

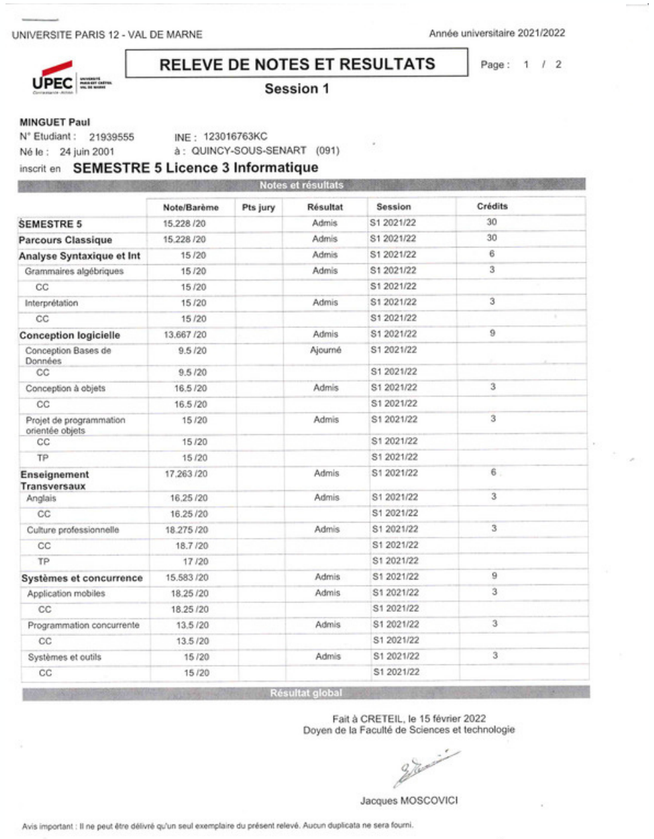
NOM	PRENOM
Hench	John
Baxter	Tony
Blair	Mary
Sherman	
Sherman	
Davis	Marc
Davis	
Sklar	Marty

Différences :

- **Le texte clair n'est pas pris en compte**

Pourcentage de similitude :
69.24%

Résultats



UNIVERSITE PARIS 12 - VAL DE MARNE — Année universitaire 2021/2022

UPEC

RELEVE DE NOTES ET RESULTATS — Page : 1 / 2

Session 1

MINGUET Paul
N° Etudiant : 21939555 — INE : 123016763KC
Né le : 24 juin 2001 — à : QUINCY-SOUS-SENART (091)
inscrit en **SEMESTRE 5 Licence 3 Informatique**

Notes et résultats

	Note/Barème	Pts jury	Résultat	Session	Crédits
SEMESTRE 5	15.228 /20		Admis	S1 2021/22	30
Parcours Classique	15.228 /20		Admis	S1 2021/22	30
Analyse Syntaxique et Int	15 /20		Admis	S1 2021/22	6
Grammaires algébriques	15 /20		Admis	S1 2021/22	3
CC	15 /20			S1 2021/22	
Interprétation	15 /20		Admis	S1 2021/22	3
CC	15 /20			S1 2021/22	
Conception logicielle	13.667 /20		Admis	S1 2021/22	9
Conception Bases de Données	9.5 /20		Ajourné	S1 2021/22	
CC	9.5 /20			S1 2021/22	
Conception à objets	16.5 /20		Admis	S1 2021/22	3
CC	16.5 /20			S1 2021/22	
Projet de programmation orientée objets	15 /20		Admis	S1 2021/22	3
CC	15 /20			S1 2021/22	
TP	15 /20			S1 2021/22	
Enseignement Transversaux	17.263 /20		Admis	S1 2021/22	6
Anglais	16.25 /20		Admis	S1 2021/22	3
CC	16.25 /20			S1 2021/22	
Culture professionnelle	18.275 /20		Admis	S1 2021/22	3
CC	18.7 /20			S1 2021/22	
TP	17 /20			S1 2021/22	
Systèmes et concurrence	15.583 /20		Admis	S1 2021/22	9
Application mobiles	18.25 /20		Admis	S1 2021/22	3
CC	18.25 /20			S1 2021/22	
Programmation concurrente	13.5 /20		Admis	S1 2021/22	3
CC	13.5 /20			S1 2021/22	
Systèmes et outils	15 /20		Admis	S1 2021/22	3
CC	15 /20			S1 2021/22	

Résultat global

Fait à CRETEIL, le 15 février 2022
Doyen de la Faculté de Sciences et technologie

Jacques MOSCOVICI

Avis important : Il ne peut être délivré qu'un seul exemplaire du présent relevé. Aucun duplicata ne sera fourni.

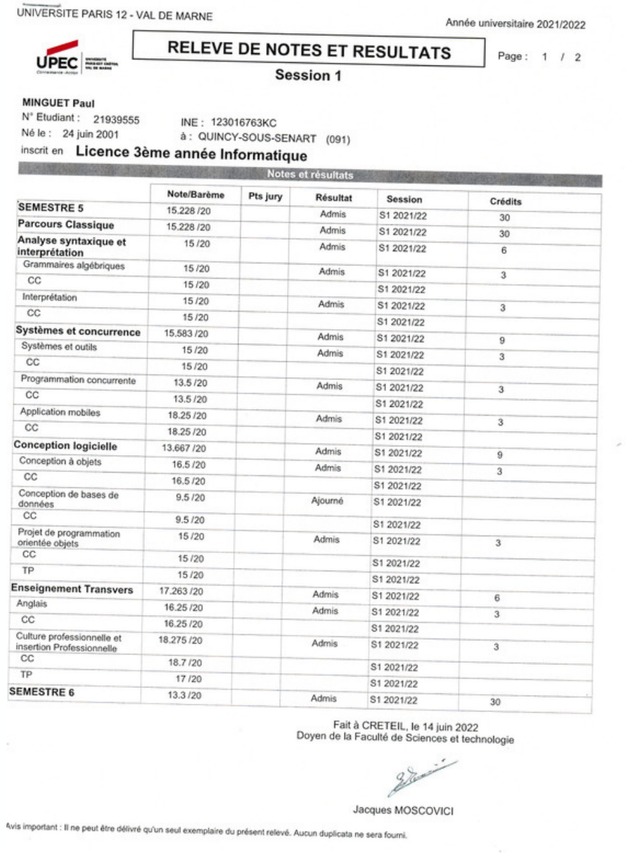
Différences :

- "S" -> "$"
- Bas de l'image rogné

Pourcentage de similitude :

77.22%

Résultats



UNIVERSITE PARIS 12 - VAL DE MARNE

Année universitaire 2021/2022

UPEC

RELEVE DE NOTES ET RESULTATS

Session 1

Page : 1 / 2

MINGUET Paul
N° Etudiant : 21939555 INE : 123016763KC
Né le : 24 juin 2001 à : QUINCY-SOUS-SENART (091)
inscrit en **Licence 3ème année Informatique**

Notes et résultats

	Note/Barème	Pts jury	Résultat	Session	Crédits
SEMESTRE 5	15.228 /20		Admis	S1 2021/22	30
Parcours Classique	15.228 /20		Admis	S1 2021/22	30
Analyse syntaxique et interprétation	15 /20		Admis	S1 2021/22	6
Grammaires algébriques	15 /20		Admis	S1 2021/22	3
CC	15 /20			S1 2021/22	
Interprétation	15 /20		Admis	S1 2021/22	3
CC	15 /20			S1 2021/22	
Systèmes et concurrence	15.583 /20		Admis	S1 2021/22	9
Systèmes et outils	15 /20		Admis	S1 2021/22	3
CC	15 /20			S1 2021/22	
Programmation concurrente	13.5 /20		Admis	S1 2021/22	3
CC	13.5 /20			S1 2021/22	
Application mobiles	18.25 /20		Admis	S1 2021/22	3
CC	18.25 /20			S1 2021/22	
Conception logicielle	13.667 /20		Admis	S1 2021/22	9
Conception à objets	16.5 /20		Admis	S1 2021/22	3
CC	16.5 /20			S1 2021/22	
Conception de bases de données	9.5 /20		Ajourné	S1 2021/22	
CC	9.5 /20			S1 2021/22	
Projet de programmation orientée objets	15 /20		Admis	S1 2021/22	3
CC	15 /20			S1 2021/22	
TP	15 /20			S1 2021/22	
Enseignement Transvers	17.263 /20		Admis	S1 2021/22	6
Anglais	16.25 /20		Admis	S1 2021/22	3
CC	16.25 /20			S1 2021/22	
Culture professionnelle et insertion Professionnelle	18.275 /20		Admis	S1 2021/22	3
CC	18.7 /20			S1 2021/22	
TP	17 /20			S1 2021/22	
SEMESTRE 6	13.3 /20		Admis	S1 2021/22	30

Fait à CRETEIL, le 14 juin 2022
Doyen de la Faculté de Sciences et technologie

Jacques MOSCOVICI

Avis important : Il ne peut être délivré qu'un seul exemplaire du présent relevé. Aucun duplicata ne sera fourni.

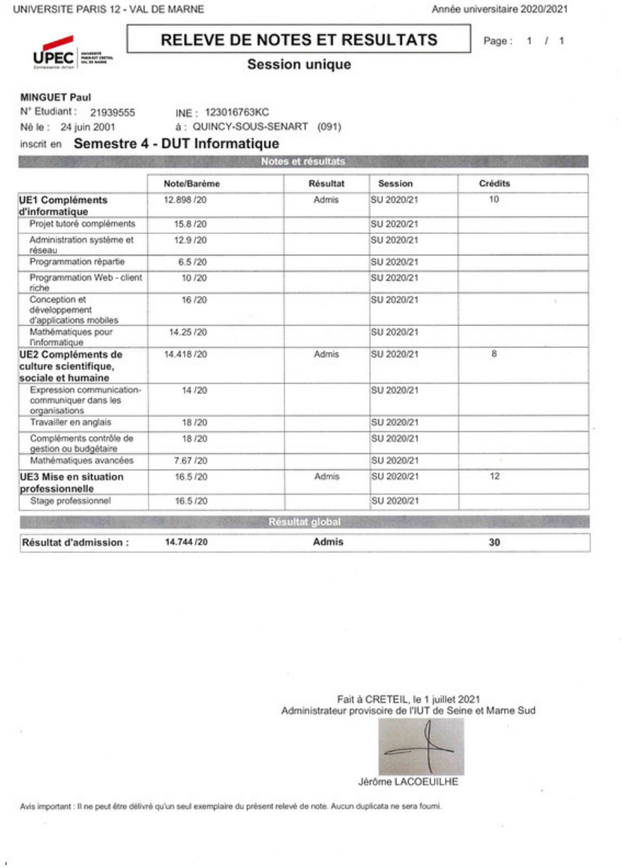
Différences :

- "S" -> "$"
- "CC" -> ""
- "TP" -> ""
- Bas de l'image rogné

Pourcentage de similitude :

75.00%

Résultats



UNIVERSITE PARIS 12 - VAL DE MARNE — Année universitaire 2020/2021

UPEC

RELEVE DE NOTES ET RESULTATS — Page : 1 / 1

Session unique

MINGUET Paul
N° Etudiant : 21939555 — INE : 123016763KC
Né le : 24 juin 2001 — à : QUINCY-SOUS-SENART (091)
inscrit en **Semestre 4 - DUT Informatique**

Notes et résultats

	Note/Barème	Résultat	Session	Crédits
UE1 Compléments d'informatique	12.898 /20	Admis	SU 2020/21	10
Projet tutoré compléments	15.8 /20		SU 2020/21	
Administration système et réseau	12.9 /20		SU 2020/21	
Programmation répartie	6.5 /20		SU 2020/21	
Programmation Web - client riche	10 /20		SU 2020/21	
Conception et développement d'applications mobiles	16 /20		SU 2020/21	
Mathématiques pour l'informatique	14.25 /20		SU 2020/21	
UE2 Compléments de culture scientifique, sociale et humaine	14.418 /20	Admis	SU 2020/21	8
Expression communication-communiquer dans les organisations	14 /20		SU 2020/21	
Travailler en anglais	18 /20		SU 2020/21	
Compléments contrôle de gestion ou budgétaire	18 /20		SU 2020/21	
Mathématiques avancées	7.67 /20		SU 2020/21	
UE3 Mise en situation professionnelle	16.5 /20	Admis	SU 2020/21	12
Stage professionnel	16.5 /20		SU 2020/21	

Résultat global

Résultat d'admission :	**14.744 /20**	**Admis**	**30**

Fait à CRETEIL, le 1 juillet 2021
Administrateur provisoire de l'IUT de Seine et Marne Sud

Jérôme LACOEUILHE

Avis important : Il ne peut être délivré qu'un seul exemplaire du présent relevé de note. Aucun duplicata ne sera fourni.

Différences :

- "tutoré" -> "tutoreé"
- Quelques erreurs de l'OCR

Pourcentage de similitude :
86.67% (81% rotation)

Résultats

Traduction de l'Arabe

République Algérienne Démocratique & Populaire

Ministère de l'Education Nationale — Office National des Examens et des Concours

6115/21
6161

RELEVE DE NOTES

Valable à l'Étranger

N° d'Inscription: 36009306

Baccalauréat Session: 2016
Filière: Sciences Expérimentales

Vu le procès-verbal de la commission des délibérations, le directeur de l'office national des examens et des concours atteste que:

Monsieur: ZEGGARI Haithem

Né le: 25/04/1998 à Constantine – Constantine

École Nationale Supérieure d'Informatique
E.S.I ex. I.N.I
Copie Conforme à l'Original

Etablissement: Lycée Sniguer Boukhemis Tamalous /6168

Est admis après avoir obtenu les notes suivantes:

Epreuves	Note/20	Coefficient	Total
Sciences de la Nature et de la Vie	17.50	6	105.00
Sciences Physiques	20.00	5	100.00
Mathématiques	18.50	5	092.50
Langue et Littérature Arabe	17.00	3	051.00
Français – Langue Etrangère 1	17.00	2	034.00
Anglais – Langue Etrangère 2	17.50	2	035.00
Philosophie	09.00	2	018.00
Histoire & Géographie	14.50	2	029.00
Langue Amazigh	-	2	--
Sciences Islamiques	17.00	2	034.00
Éducation Physique et Sportive	19.50	1	19.50
Total Général:		30	518.00

Moyenne Générale /20 : 17.27

N° secret
JPSSCPhF

Fait à Alger, le: 11/07/2016

Le Directeur de l'Office National des Examens et des Concours
M. BENZEMRANE
Signature Illisible

haithem.arcelona/1007USTPWD_BC---1468241011

Suit l'empreinte d'un sceau officiel rond portant les mentions :
• République Algérienne Démocratique et Populaire
Ministère de l'Education Nationale
Office National des Examens et des Concours
Le Directeur

Nota : Il n'est délivré qu'un seul exemplaire du présent relevé.

Pour traduction conforme

08 DEC. 2020

Scanné avec CamScanner

Différences :

- Quelques erreurs de l'OCR

Pourcentage de similitude :

76.52%

Résultats

ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE
Direction des études

الجمهورية الجزائرية الديمقراطية الشعبية
République Algérienne Démocratique et Populaire
وزارة التعليم العالي و البحث العلمي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

المدرسة الوطنية العليا للإعلام الآلي
Ecole nationale Supérieure d'Informatique
مديرية الدراسات

RELEVE DE NOTES

Matricule : **16/0116** Année universitaire : **2018/2019** Année d'étude : **1ère année - Second Cycle**

Nom : **ZEGGARI** Prénom : **HAITHEM** Date de naissance : **25/04/1998** à : **CONSTANTINE**

Diplôme préparé : **Ingénieur d'état en Informatique**

RES1	Réseaux 1	4	10.65	Février
SYS1	Systèmes d'exploitation 1	5	5.50	Février
IGL	Introduction au génie logiciel	5	10.00	Février
THP	Théorie des langages de programmation et applications	4	11.90	Février
ANUM	Analyse Numérique	4	11.50	Février
ORG	Analyse des organisations	3	14.30	Février
RO	Recherche Opérationnelle: graphes et algorithmes	3	7.20	Février
LANG1	Langue anglaise 1	2	12.17	Février
ARCH	Architecture	4	13.33	Juin
RES2	Réseaux 2	3	12.11	Juin
SYS2	Système d'exploitation 2	4	8.35	Juin
BDD	Bases de données	5	12.00	Juin
MCSI	Méthodologies d'analyse et conception de systèmes d'information	5	12.80	Juin
CPROJ	Conduite de projets	3	10.88	Juin
PROJ	Projet	3	15.51	Juin
SEC	Introduction à la sécurité informatique	1	15.00	Juin
LANG2	Langue anglaise 2	2	14.50	Juin

Moyenne Annuelle: **11.21/20**

Fait le: **25/01/2021** Décision du Conseil : **Admis** Rang : **130** sur **229**

رئيس قسم
متخصصة المعلومات والتنمية لحاسوب
بالمدرسة الوطنية العليا للإعلام الآلي

Il n'est délivré qu'un seul exemplaire de ce document. لا تسلم الا نسخة واحدة من هذه الوثيقة

Ecole nationale Supérieure d'Informatique المدرسة الوطنية العليا للإعلام الآلي BP68 16270, Oued Smar, Alger. Tél : 023939132 ; Fax : 023939142 ; http://www.esi.dz

Différences :

- Quelques erreurs de l'OCR

Pourcentage de similitude :

92.43%

Résultats

Caractéristiques de l'image	Pourcentage moyen	Principales erreurs
Simple, texte avec symboles ([,(;:)])	88.08%	Certains caractères non / mal reconnus
Texte en couleur	96.25%	
Trous dans le tableau	92.66%	
Trous et couleurs	84.56%	
Couleurs claires	56.04%	Certaines couleurs disparaissent au pré-traitement

Résultats

Trous et couleurs claires	70.52%	Certaines couleurs disparaissent au pré-traitement
Images scannées	88.04%	Certains caractères non / mal reconnus et certaines images coupées
Tableaux sans bordure	0.00%	
Tableaux tournés	Idems que non tournés	

Conclusion

Ce qu'on a appris :

- **Pré-traitement des images**
- **Techniques de recadrage d'images**
- **Implémentation d'un logiciel d'OCR dans un script**

Version 2.0 :

- **Amélioration du pré traitement**
- **Images sans bordures / grille**

Pour recommencer :

- **Plus de tests sur le pré traitement des images**
- **Plus de recherche sur la technologie**

Merci pour votre attention

Avez-vous des questions ?