Projet

PenPyText

24/01/2022

DataScientest

LA TEAM PenPyText

SAMY AIT AMEUR

SOPHIE AMEDRO

STEPHANE TCHATAT

DataScientest

LE PROJET PenPyText EN SYNTHÈSE

01

PROBLÉMATIQUE & ENJEUX

02

LES DATAS

03

EXPLORATION & PRÉPARATION

04

MODÈLES UTILISÉS

05

ANALYSE

06

CONCLUSION & PERSPECTIVES

DataScientest

PROBLÉMATIQUE & ENJEUX

DataScientest

Reconnaître pour/et digitaliser des textes manuscrits

Nécessité de numériser des documents manuscrits :

- Faciliter et organiser le stockage
- Pérenniser/sécuriser l'archivage

Reconnaître pour/et digitaliser des textes manuscrits

Nécessité de numériser des documents manuscrits :

➡ Faciliter et organiser le stockage

➡ Pérenniser/sécuriser l'archivage

- Exploitation d'un **jeu de données** important, comportant des erreurs ou malformations
- Mise en œuvre de différentes **techniques de pré-traitement** des images
- Construction d'un **réseau de neurones** avec les contraintes techniques du jeu de données, mais aussi de la problématique abordée
- **Interprétation et analyse** des résultats intermédiaires et finaux
- **Restitution d'un modèle** et présentation des résultats

LES DATAS

DataScientest

Les données

Les données

Utilisation du jeu de données IAM Handwriting Database 3.0

- Formes de texte anglais
- 1 539 pages de texte de natures variées
- 657 personnes ont participé
- Ecritures manuscrites sans contrainte
- Scannées à une résolution de 300 dpi
- Images PNG avec 256 niveaux de gris

- 5 685 phrases isolées et labellisées
- 13 353 lignes de texte isolées et labellisées

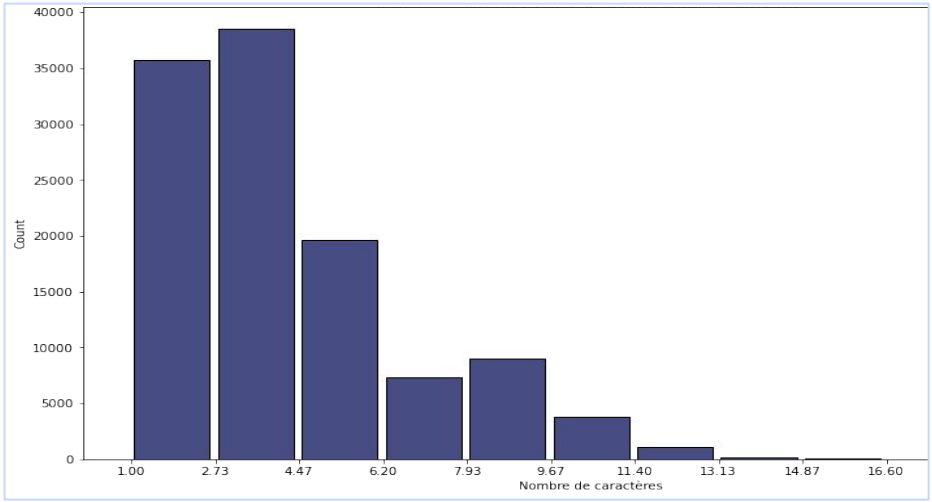
➡ 115 320 mots isolés et labellisés

EXPLORATION & PRÉPARATION

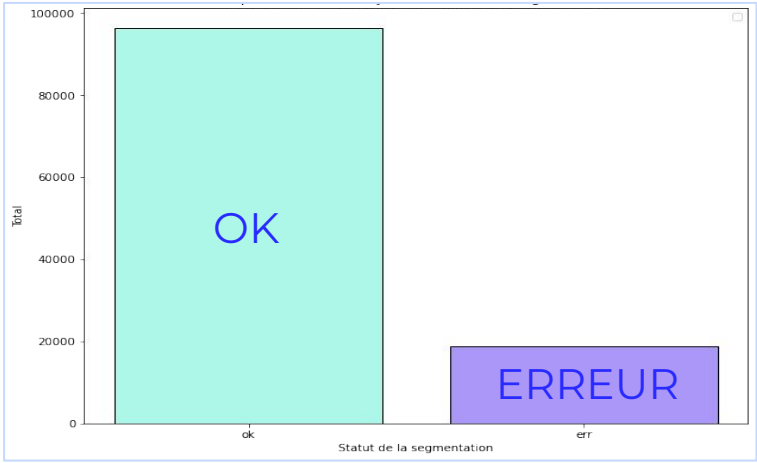
DataScientest

Distribution du nombre de caractères

Distribution du nombre de caractères par transcription

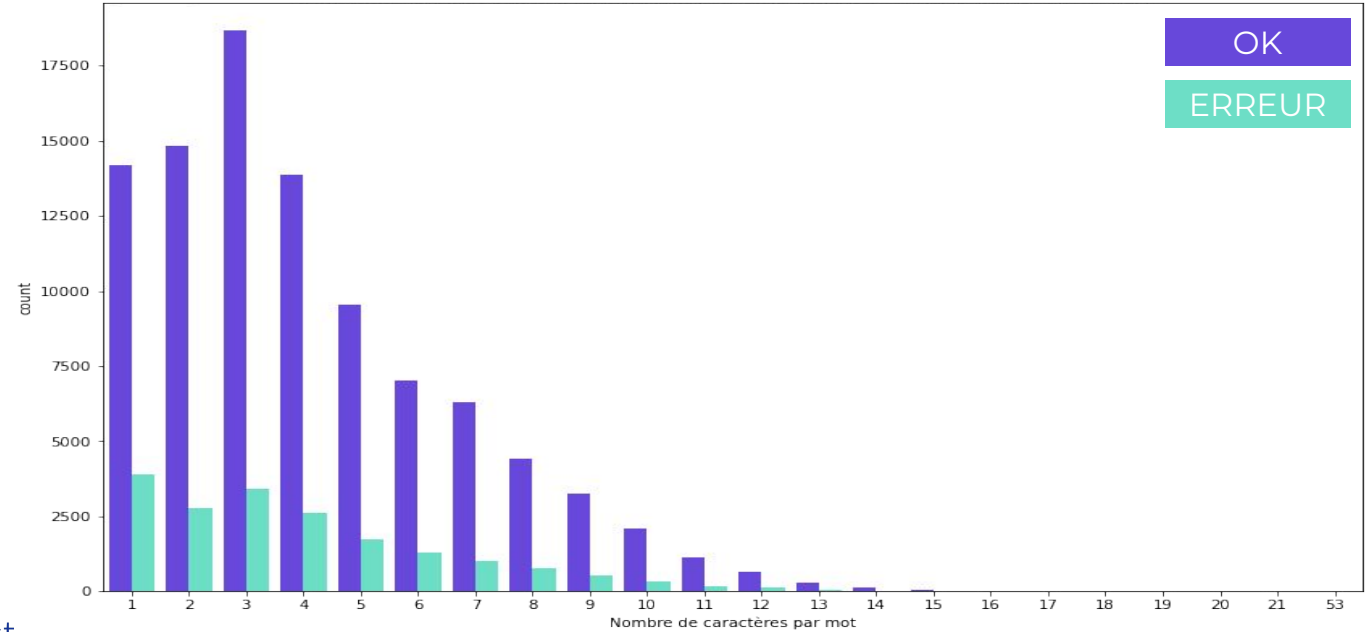


Proportion des mots
ayant une erreur de segmentation



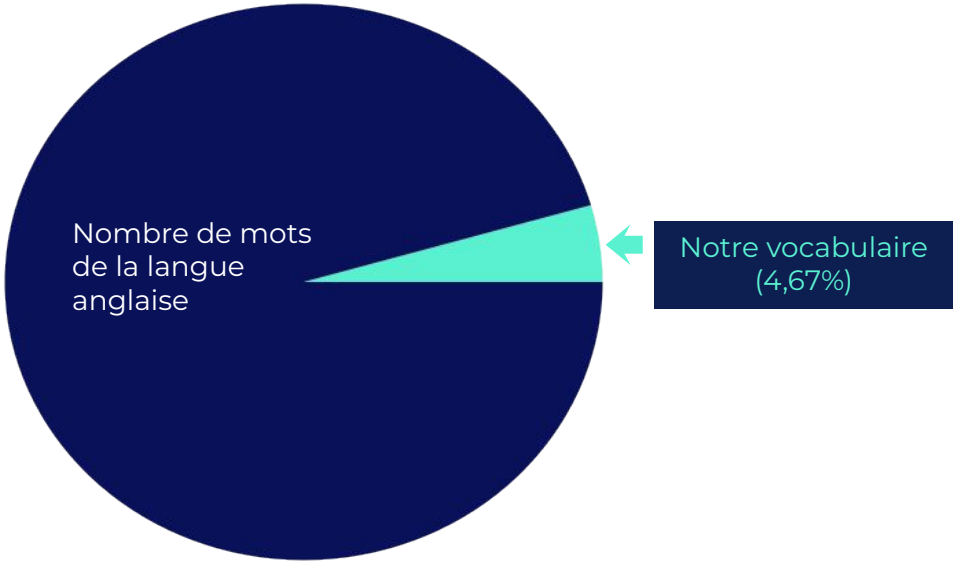
DataScientest

Distribution du nombre de caractères par transcription suivant le statut de segmentation



DataScientest

Selon l'Oxford English Dictionary, 171 476 mots sont couramment employés à l'heure actuelle



DataScientest

Pipeline de préparation des données

- Transformation des données en dataframe
- Transformation des images

  Retrait du jeu de données les images endommagées

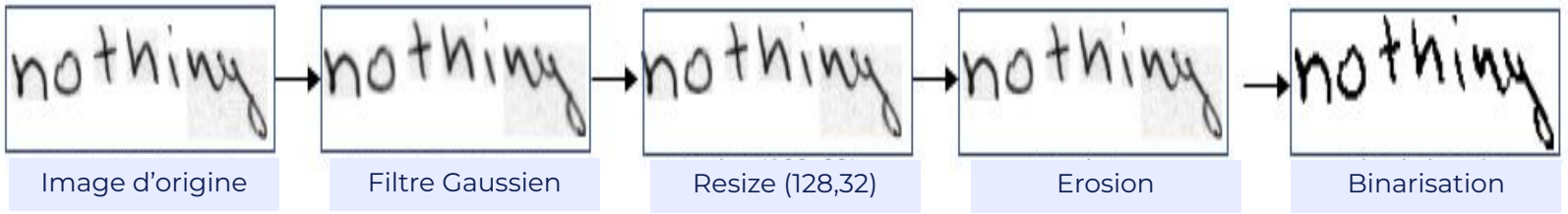
  Nettoyage du bruit

  Harmonisation de la taille des images

  Érosion

  Passage des images en noir et blanc

  Retrait du trainset les images avec erreur de segmentation



MODÈLES UTILISÉS

DataScientest

Choix du modèle et optimisation

Traitement de 2 approches des réseaux de neurones pour la classification

- Choix de mots entiers comme classes
- Obtention d'un dictionnaire de 8004 mots

➡ Modèle de réseau de neurones convolutifs pour extraire les caractéristiques des images et obtenir une classification

- Reconnaissance de caractères

➡ Modèle de réseau de neurones récurrents associé à un algorithme de rétro-propagation

TABLEAU DE COMPARAISON DES CNN

LAYERS	TEMPS PAR EPOCHS	ACCURACY
Conv(k=(9,9,s=(2,2),64) Conv(k=(5,5),128) Conv(k=(3,3),128) Conv(k=(3,3),256) Conv(k=(3,3),256) Dense(512)	**215 s**	**pred 1: 0.224852 top_5: 0.596446**
type vgg16* Conv(k=(3,3),64) MaxPool((2,2), s=(2,2)) Conv(k=(3,3),128) *2 MaxPool((2,2), s=(2,2)) Conv(k=(3,3),256) *3 MaxPool((2,2), s=(2,2)) Conv(k=(3,3),512) *3 MaxPool((2,2), s=(2,2)) Conv(k=(3,3),512) *3 Dense(4096) Dense(4096)	300 s	pred 1:0.21787 top_5: 0.599569
type lenet* Conv(k=(5,5),32) MaxPool((2,2)) Conv(k=(3,3),64) MaxPool((2,2)) GlobalAveragePooling Dense(512) Dense(1024)	35 s	pred 1:0.086796 top_5: 0.315622

DataScientest

TABLEAU DE COMPARAISON DES CRNN

LAYERS	MAX_LEN	TEMPS PAR EPOCHS	ACCURACY	CER MOYENNE SUR PRED 1
Conv(k=(3,3),64) MaxPool((2,2)) Conv(k=(3,3)128) MaxPool((2,2)) Conv(k=(3,3),256) MaxPool((2,2)) Dense(256) BLSTM(128) BLSTM(64)	10	760 s	top: 0.395833 top_5: 0.495453	0.4001
Conv(k=(3,3),64) MaxPool((2,2)) Conv(k=(3,3)128) MaxPool((2,2)) Conv(k=(3,3),256) MaxPool((2,2)) Dense(256) **GRU**(128) **GRU**(64)	10	780 s	top: 0.414192 top_5: 0.51715	0.3780
Conv(k=(9,9),s=(2,2),64) Conv(k=(5,5)128) MaxPool((2,2)) Conv(k=(3,3),256) MaxPool((2,2)) Dense(256) Dense(128) BLSTM(64) BLSTM(64)	10	760 s	top: 0.432666 top_5: 0.52866	0.3653

DataScientest

ANALYSE

DataScientest

Analyse des erreurs et améliorations

Modèle CNN

- Des mots absents de notre dictionnaire d'origine
- Une écriture manuscrite dont les caractéristiques n'ont pas été répertoriées, ne peut pas être transcrite correctement
- Nécessité d'avoir un jeu de données plus important avec une augmentation de la taille de notre vocabulaire

Choix du modèle avec RÉSEAUX DE NEURONES RÉCURRENTS

- 1ère amélioration : première prédiction, accuracy de 40% avec un modèle CRNN (versus 20% avec un modèle CNN)
- 2ème amélioration : plus de limite de vocabulaire. Tout mot peut être transcrit

Analyse des erreurs et améliorations

Modèle CRNN

- 1ers entraînements : progression des résultats en affinant la partie CNN du modèle
- Mots de plus de 4 lettres n'étaient pas prédits
- Plus faible représentation des mots plus longs dans le dataset
- Choix de se concentrer sur des mots de maximum 10 caractères
- Amélioration d'environ 10% de notre accuracy sur les cinq premières prédictions
- Persistance d'erreurs sur mots de moins de 5 caractères dues à des images mal segmentées

Recours aux ressources de Stackoverflow et exemples de code disponibles sur Github pour résoudre des problèmes d'implémentation

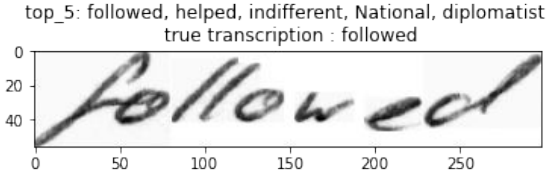
- Retrait des images mal segmentées du jeu de données d'entraînement
- Amélioration de l'accuracy de l'ordre de 5%
- Réduction de la taille du lexique de caractère, en supprimant les caractères non présents dans notre jeu de données : gain en en accuracy

Analyse du meilleur modèle

Evaluation CNN

Première prédiction
- 0.775148 faux
- 0.224852 vraies

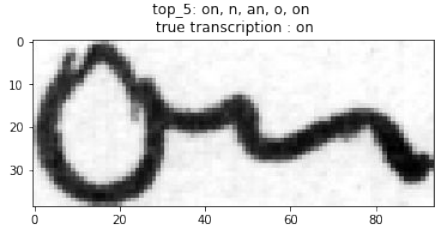
5 premières prédictions
- 0.596446 vraies
- 0.403554 faux



Evaluation RNN

Première prédiction
- 0.567334 faux
- 0.432666 vraies

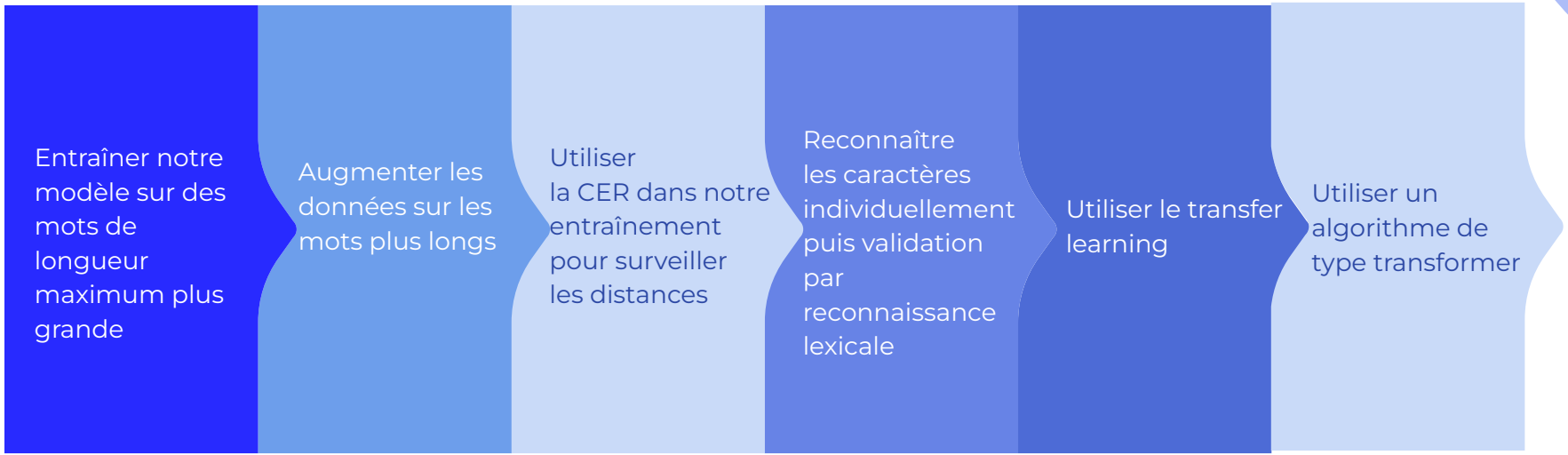
5 premières prédictions
- 0.52866 vraies
- 0.47134 faux



CONCLUSION & PERSPECTIVES

DataScientest

Pour aller plus loin...



Pour conclure...

Un projet synonyme de challenge
et d'apprentissage en compétences

Des résultats satisfaisants
Programmation de notre propre générateur de données pour le CRNN
Accuracy correcte pour un premier projet d'OCR

Un projet qui nous conforte dans notre choix d'explorer
professionnellement la galaxie des Datas sciences