Création de données

Cours M1 TAL - Inalco

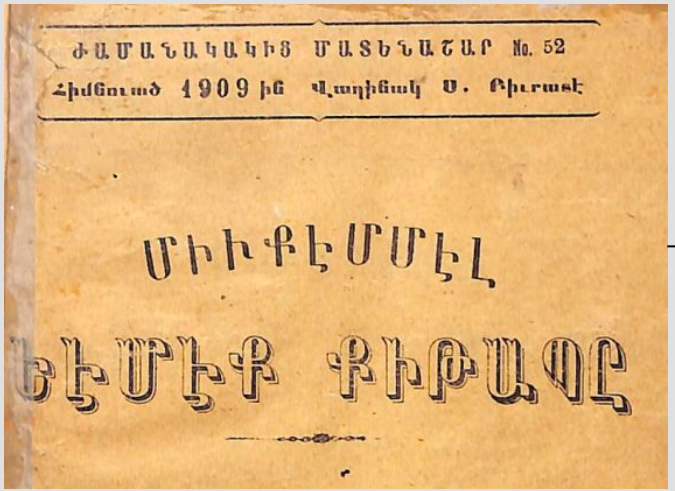
Johanna Cordova

Création d’un corpus textuel

Numérisation et structuration

Numérisation de textes : OCR

La reconnaissance optique de caractères (abrégé en OCR pour *optical text recognition*) est la tâche qui consiste à obtenir, à partir d'une image, le texte imprimé contenu dans celle-ci.



ժա-ԺԱՆԱԿԱԿԻՑՏ ՄԲՐՄԱՏԵՖԱՇԱՐ N52
Հիմնուած 1909ին ՎԱՂԻՆԱԿ Ս. ԲԻՒՐԱՏԷ
ՄՍԻԻՐԷՄՄԷԼ
ԵԷՄԷՔ ֆԲԻԹԱՂՊԸ

Numérisation de textes : HTR

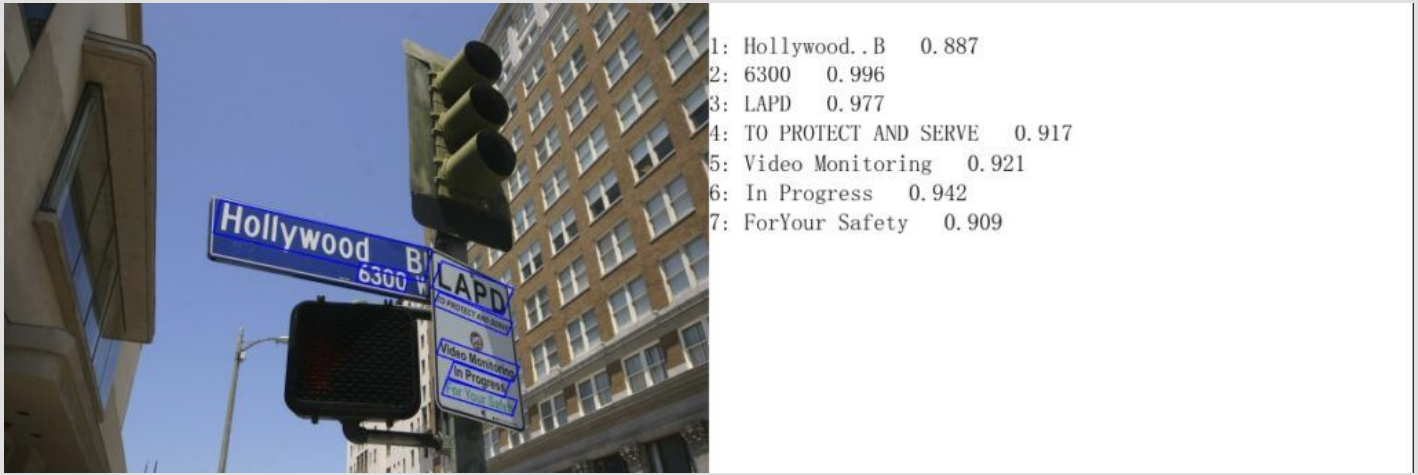
La reconnaissance de l'écriture manuscrite (*handwritten text recognition*, HTR) consiste de même à extraire du texte manuscrit à partir d'une image. Cette tâche est plus complexe : la forme des caractères varie fortement d'un scripteur à l'autre.

Manuel de turc en arabe, 1837, archives de la Bulac. Source : Gallica

Détection de texte dans des scènes (STR)

Une autre tâche consiste à détecter et reconnaître du texte dans des scènes complexes et non structurées (*Scene Text Recognition*) et non plus sur un support uniforme.

Cette tâche est traitée par des modèles dont l'architecture diffèrent de l'OCR.



Principe de la reconnaissance de caractères

La reconnaissance de caractère s'effectue en trois étapes :

1) un **prétraitement** de l'image (conversion en niveau de gris) pour distinguer le texte du fond ;
2) la **segmentation** du texte pour identifier les blocs de texte, les lignes, les mots puis les caractères ;
3) la **reconnaissance des caractères** .

Outils d'OCR multilingues

- **Tesseract** : outil open-source, modèles pour 127 langues et 37 scripts. S'utilise en ligne de commande ou à travers une librairie Python (pytesseract).
- **Paddle** : outil de STR et d'OCR, permet de produire des documents structurés (en JSON ou Markdown par exemple). Couvre + de 80 langues.
- **EasyOCR :** outil de STR et d'OCR, couvre + de 80 langues. S'utilise à travers une librairie Python.

- **Kraken** : développé dans un cadre académique, dispose de modèles spécialement entraînés pour la recherche en sciences humaines

Outils d'HTR

Pour faire de l'HTR, il est souvent nécessaire d'entraîner ses propres modèles pour les adapter à l'écriture particulière qu'on veut reconnaître.

Principaux outils d'HTR pour la recherche (étude de manuscrits historiques) :

- **Transkribus**
- **eScriptorium**

Quel outil choisir ? Calculer la fiabilité

La fiabilité d'un outil de reconnaissance de texte se mesure grâce à 2 métriques :

- Le taux d'erreur par mot (*Word Error Rate*, **WER**)
- Le taux d'erreur par caractère (*Character Error Rate*, **CER**)

On comptabilise 3 types d'erreurs, qui déterminent la distance d'édition :

- **Insertion** : caractère en trop
- **Délétion** : caractère manquant
- **Substitution** : un caractère à la place d'un autre

Calculer la fiabilité

CER = (I + D + S) / nombre total de caractères de la référence

WER = (I + D + S) / nombre total de mots de la référence

RÉFÉRENCE (vérité-terrain)

Le lendemain matin, Catherine peigna les grands cheveux noirs de sa petite fille avec soin

SORTIE DE L'OCR (hypothèse)

Le lndemain matin, Catherino peigna les grands cheveux noirs des sn petite fillo avec soin

Exercice : utiliser Tesseract

Installation de Tesseract 5

Sur Linux :

```
sudo apt install tesseract-ocr
sudo snap install --channel=edge tesseract
```

Sur MacOS :

```
brew install tesseract
```

Installation de Tesseract 5

Les modèles pour les différentes langues peuvent être téléchargés à part :

https://github.com/tesseract-ocr/tessdata_best

Ils sont identifié par le code ISO de la langue qu'ils couvrent.

Par défaut, ces modèles sont stockés :

- Sur Linux : `/usr/share/tesseract-ocr/5/tessdata/`
- Sur Mac : `/opt/homebrew/Cellar/tesseract-ocr/5/tessdata`

(sur Mac, exécuter la commande `brew list tesseract` pour trouver le chemin exact)

Importer des modèles

Aller sur : https://github.com/tesseract-ocr/tessdata_best

- Télécharger le modèle dont vous avez besoin (l'extension des modèles est `.traineddata`), puis identifier le chemin d'accès de ce modèle sur votre ordinateur. Par exemple : `/home/johanna/Documents/Cours/Création_de_données/hin.traineddata`
- Ouvrir un terminal et se placer dans le répertoire `tessdata` identifié précédemment :

  Linux : `cd /usr/share/tesseract-ocr/5/tessadata`

  Mac : `cd /opt/homebrew/Cellar/tesseract-ocr/5/tessdata`

- Déplacer ensuite ce modèle dans le répertoire avec la commande suivante :

```
sudo mv /home/johanna/Documents/Cours/Création_de_données/hin.traineddata .
```

Utilisation de Tesseract

Se placer dans le répertoire contenant l'image à transcrire et ouvrir un terminal.
L'image doit être au format PNG ou au format TIFF.

La commande suivante permet de lancer la reconnaissance (remplacer ISO par le code de la langue concernée) :

```
tesseract -l ISO image.png image
```

Si le texte contient plusieurs langues, indiquer toutes les langues :

```
tesseract -l fra|eng image.png image
```

Structurer le corpus : XML-TEI

XML-TEI

Le **format XML** (*eXtensible Markup Language)* permet de structurer les informations d'une page et d'ajouter des métadonnées au contenu.

Le **schéma TEI** (*Text Encoding Initiative*) est une proposition de convention de balisage XML destinée à structurer des textes (littéraires, historiques) pour l'édition numérique. Très utilisé en Humanités Numériques où les métadonnées internes aux textes sont essentielles.