1 La reconnaissance du texte et des segments

Le premier étape du pipeline *Gallic(orpor)a* est la reconnaissance du texte dans les images numérisées. Ces images se composent des pixels, chacun d'eux se compose d'un tableau de numéros pour décrire le degré de rouge, bleu, et vert dans le carré. Pour arriver donc d'un rassemblement de pixels au caractère d'un système d'écriture, il faut un modèle HTR qui sait chercher dans les pixels les configurations des caractères. Avec la reconnaissance de texte, il faut un deuxième modèle qui sait reconnaître les régions cohérentes sur la page. Ce dernier modèle cherchent aussi dans les pixels pour les configurations consistantes, mais au lieu de reconnaître dedans des caractères, il relève les polygons ou les rectangles qui contient une entité cohérente.

1.1 La création des données d'entraînement

1.2 L'entraînement des modèles

1.3 Les modèles prédisent le texte

2 La reconstitution des données

3 L'analyse linguistique

3.1 La création des données d'entraînement

3.2 L'entraînement des modèles

3.3 Les modèles analysent le texte prédit

4 Le texte pré-édité