Numérisation d’œuvres littéraires

Problématique : Le Partage d’une ressource, la littérature

Comment partager les œuvres papiers ?

* Plus que partager, les numériser pour les traiter
* Mettre en place un algorithme qui transforme un texte imprimé sur du papier en texte numérique

La reconnaissance de caractères : un problème vaste

* On a besoin d’une reconnaissance de caractères particulière
  + Texte d’imprimerie
  + La police de caractère est connue
  + Pas de déformation de texte (prise de photo optimale)
* Seule la reconnaissance de caractères est abordée
* Prise de photo supposée optimale

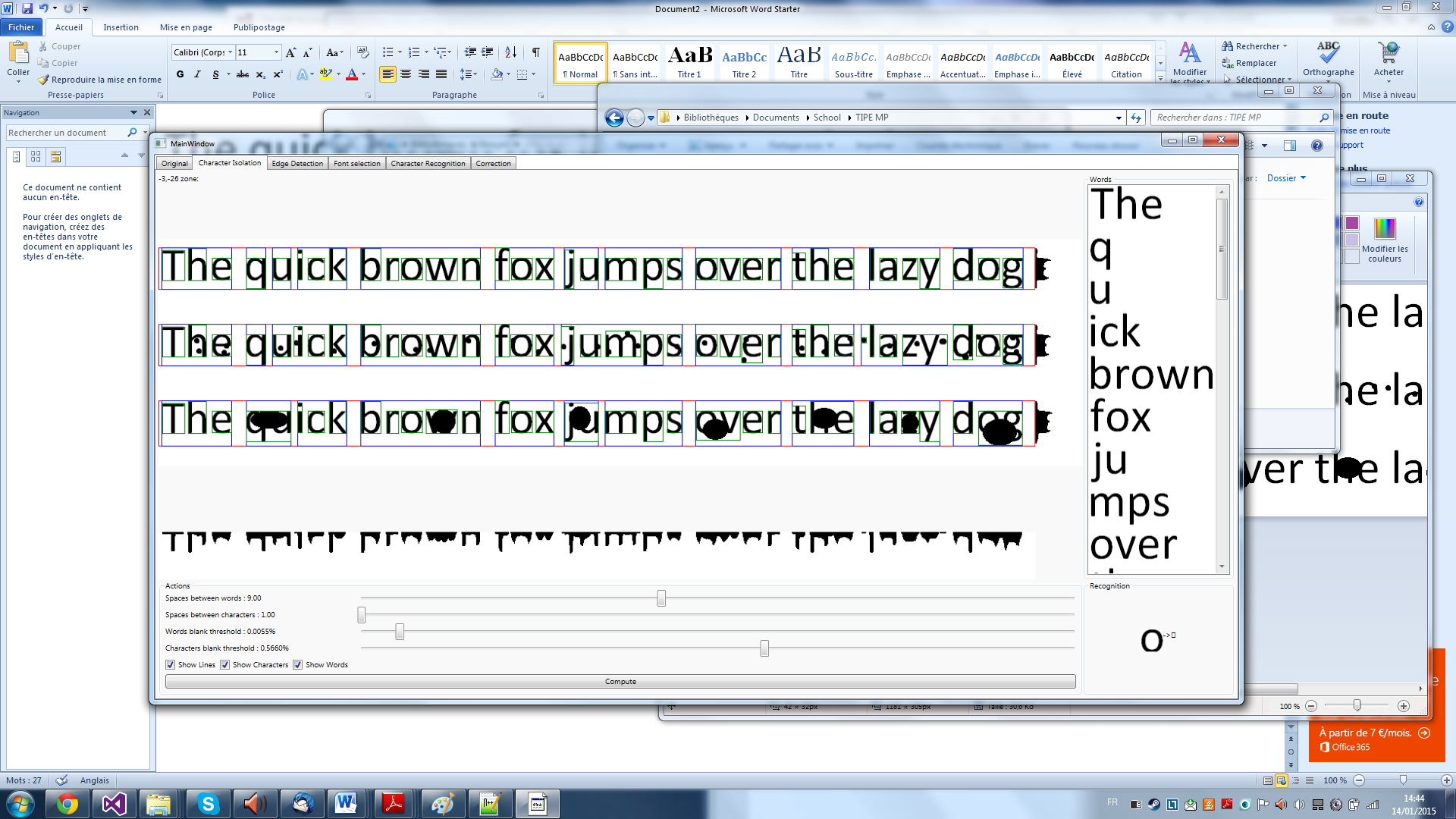
Les étapes de l’algorithme

# Isolation des caractères

# Reconnaissance des caractères

# Correction de la reconnaissance

Isolation des caractères

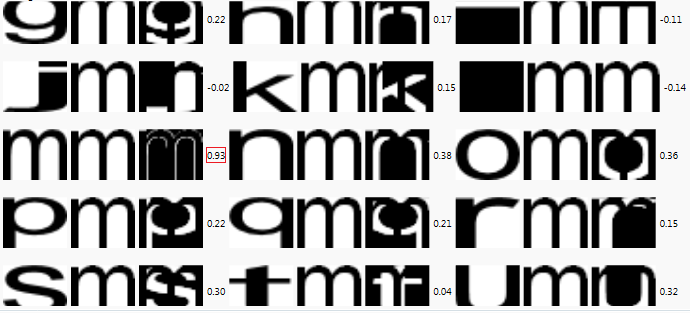


1. Isolation des lignes de textes : parcourt de l’image ligne par ligne
2. Isolation des mots : parcourt de chaque ligne
3. Isolation des caractères : parcourt de chaque mot

* Problème de décalage rencontré, lettres collés peu discernables …

Reconnaissance des caractères

* Comparaison pixel par pixel avec les caractères de références



* Seuil de confiance : proportions \* ratio de pixels corrects

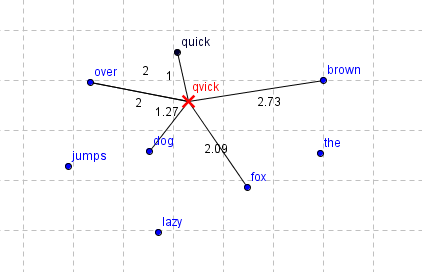
Correction de la reconnaissance

Associer l’ensemble des mots à un espace muni d’une distance

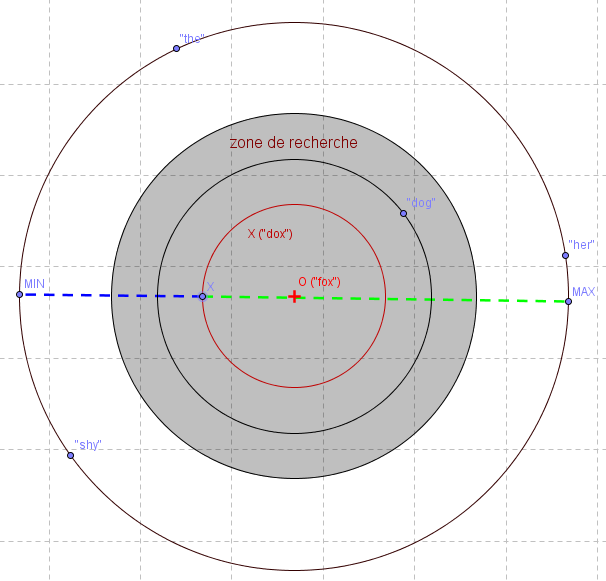
* Distance de Hamming : comparaison des caractères un à un

|  |  |  |  |
| --- | --- | --- | --- |
| D | O | G | * 2 |
| F | O | X |

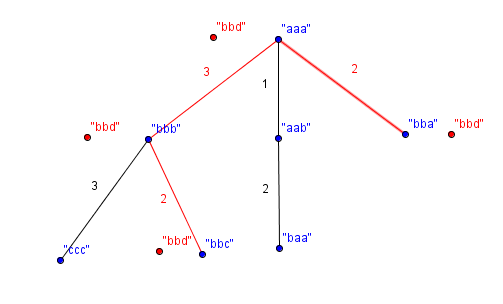
* Trouver le mot le plus proche



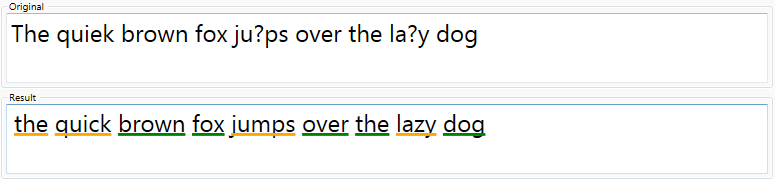
* Comparaisons couteuses en temps : optimisation par les arbres











Conclusion

* Mise au point d’un algorithme de reconnaissance de caractères
* Les œuvres ainsi numérisés sont prises en compte par les moteurs de recherches (ex : au CDI)

Nombreuses améliorations possibles

* + Reconnaissance basé sur les lignes de construction d’un caractère
  + Séparation de caractères collés
  + Recoller des mots mal découpés
  + …