# CORNELIUS – Principes algorithmiques

## Extraction du texte

*Compléter la doc.*

## Correction des lignes

Si dans un bloc d’une même ligne, on détecte un espace important entre deux blocs de mots, on crée deux blocs (lignes) distincts.

S’il existe une intersection entre deux lignes, on fusionne les deux lignes.

## Détection du sommaire

En « béta ». Utilisé pour l’instant.

## Elimination des lignes redondantes

Constitution de la liste des lignes présentes dans la partie haute ou partie basse de la page (1/5ième de la hauteur de page).

Pour chacune de ces lignes, on cherche s’il existe d’autres lignes avec un texte « similaire » (algorithme de distance de Levenshtein).

On regroupe ensuite les lignes similaires trouvées selon leurs positions sur la page.

Si un couple « ligne similaire / position sur la page » est présent sur au moins 1/5ième des pages, toutes les lignes situées au-dessus (dans le cas où le couple est en partie haute de la page) ou au-dessous (dans le cas où le couple est en partie basse de la page) sont ignorées pour le reste du traitement.

## Regroupement des lignes en blocs

Regroupement des lignes en blocs de lignes en fonction de la distance qui les sépare.

## Recherche des titres et sous-titres

On considère que le style principal du document, c’est-à-dire le style du texte, est celui qui compte le plus de caractères correspondants.

Pour chaque style ayant une taille de police supérieure à la police du style principal, on évalue sa probabilité pour que celui-ci soit un style rattaché à un titre.

Pour chaque bloc de texte qui corresponde au style :

* Si le bloc contient plus de trois lignes, la probabilité de style rattaché à un titre diminue.
* Si le bloc contient plus de deux styles, la probabilité de style rattaché à un titre diminue.
* Si le rapport nombre de caractères/nombre de ligne est trop faible, on ignore le style.
* Si le texte du bloc matche avec les expressions régulières définissant un titre, la probabilité de style rattaché à un titre augmente :
	+ 11 Titre
	+ 11.22 Titre
	+ 11.22.33 Titre
	+ 11.22.33.44 Titre
	+ 11. Titre
	+ 11.22. Titre
	+ 11.22.33. Titre
	+ 11.22.33.44. Titre
	+ A Titre
	+ A. Titre

Au final, si la probabilité que le style soit rattaché à un titre est suffisante, on flag le style comme candidat.

Si aucun style n’est candidat, on relance la procédure sans tester le match avec les expressions régulières.

Chaque bloc correspondant à un style candidat est marqué comme bloc de « rupture » (titre).

Pour chaque style candidat, on essaye de déterminer la « profondeur » du titre en analysant le texte de chaque bloc correspondant au style.

## Recherche des légendes

Pour chaque image du document, on recherche les blocs les plus proches de l’image.

On extrait les styles des blocs trouvés et on les ordonne par taille de police décroissante.

Si on trouve un style qui a une taille inférieure ou égale à la taille du style par défaut + 2 points, ce style est marque comme « légende ».

## Regroupement par section

Une section de niveau 0 est créée par défaut.

On parcourt chaque bloc de chaque page.

Si le bloc n’est pas de style « rupture », on le rattache à la section courante.

Si le bloc est de style « rupture », on crée une nouvelle section qui devient la section courante. On parcourt ensuite les sections en « remontant » afin de déterminer à quelle section la nouvelle section doit être rattachée.