Comment nettoyer Sunstein (le bougre)

Les PDFs sont sur le github dans `data/sunstein/pdf/` et les fichiers texte dans `data/sunstein/processed/`

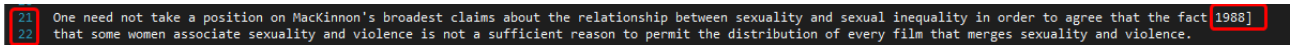
Pour chaque article, partir soit du fichier texte `cermine`, soit du fichier texte `lapdf`. En général, je pars du `lapdf`, mais si vous voyez qu'il a des gros problèmes (notamment pour les articles qui ont en deuxième page une page énorme avec écrit CHICAGO et un logo en-dessous), prenez l'autre. Copiez le fichier avant de l'éditer, et la version éditée doit avoir exactement le même nom, mais avec `lapdf` ou `cermine` remplacé par `mclean`. Ouvrez l'article original en PDF à côté. Je recommande d'utiliser Notepad++ comme éditeur de texte

Note : Il y a pas besoin de lire tout le texte, notamment pour les deux derniers points de la liste, corrigez que les erreurs que vous voyez, le but c'est pas de passer des heures sur chaque article non plus.
Note 2 : Il y a des articles qui prennent très peu de temps à nettoyer, et d'autres qui peuvent prendre une heure...

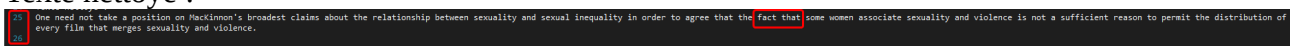
Choses principales à faire :
- Vérifier chaque transition de page. Les phrases sont généralement coupées en deux à cet endroit donc il faut supprimer le saut de ligne. Attention, parfois il y a une ou plusieurs pages qui manquent du fichier texte donc il faut bien vérifier toutes les transitions une par une. Si y a des pages qui manquent, copiez-collez le texte du pdf à la place.
- Vérifier les sauts de ligne. Le but c'est qu'une ligne = un paragraphe (ou un titre de section). Il y a des endroits où l'OCR a mis une ligne de texte pour chaque ligne du pdf, donc dans ce cas il faut reconstituer les paragraphes en supprimant les sauts de ligne et en reconstituant les mots s'il y a des césures (à faire à la main! Certaines des césures sont normales si les mots contiennent naturellement un tiret, "18-year-old" est correct mais "constitu-tion" clairement pas).
- Supprimer la majorité des numéros de référence dans le texte (c'est pas grave si vous en oubliez. Ils se manifestent souvent par des chiffres à la fin des mots ou de phrases, ou des fois l'OCR les remplace des guillemets (par ex. "...in punitive damage awards.**4** This investigation..."
- Si les notes de bas de page apparaissent au milieu du texte (à la fin de chaque page), les virer
- S'il y a le temps, copier-coller les notes de bas de page à la fin du document, reconstituer les lignes correctement (une ligne par note, même s'il y a plusieurs paragraphes), et supprimer leurs numéros. Avec Notepad++ vous pouvez sélectionner en appuyant sur Ctrl pour enlever les premiers caractères de chaque ligne par exemple. Attention, il y a souvent pas mal de texte à corriger dans les notes de bas de page, vu que le texte est plus petit et l'OCR plus mauvais
- Si vous voyez des tirets entre deux mots qui servent à introduire une proposition, rajoutez un espace des deux côtés ("well-defined" → "well-defined", mais "I think-and this is true-that I like dogs." → "I think - and this is true - that I like dogs."). Si c'est des tirets longs (—) comme c'est le cas dans certains articles, rien besoin de corriger.
- S'il y a des erreurs d'OCR qui sautent aux yeux dans le texte, corriger les mots en question.
- Pour les articles avec en deuxième page CHICAGO + le logo énorme, l'OCR remplace les chiffres et les minuscules dans les titres par des caractères "boîte avec point d'interrogation (□)" ou équivalent, c'est une erreur d'encodage, laissez-les comme ça, je les corrigerai automatiquement à la fin.

EXEMPLE 1 (TRANSITION DE PAGE)

Texte original :

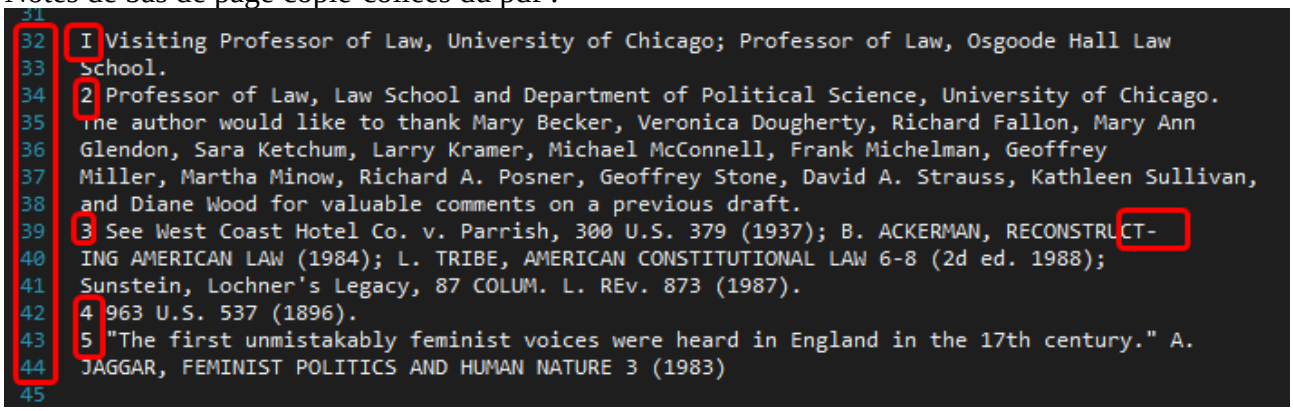


Texte nettoyé :

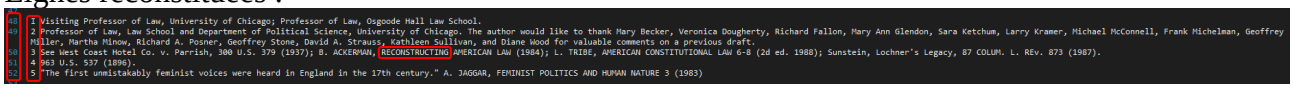


EXEMPLE 2 (NOTES DE BAS DE PAGE)

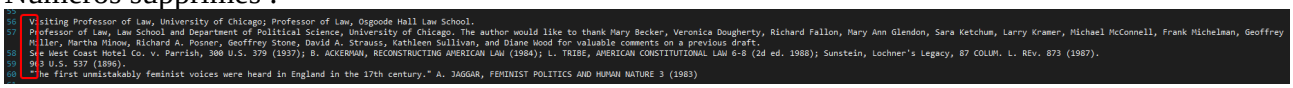
Notes de bas de page copié-collées du pdf :



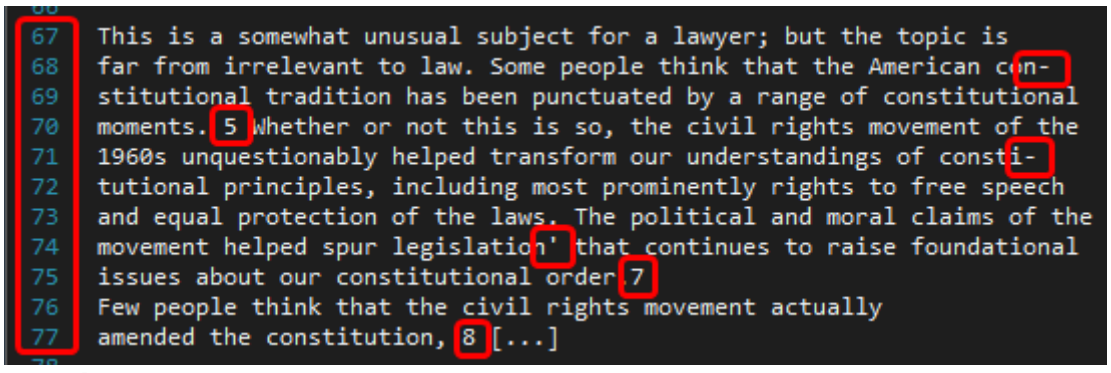
Lignes reconstituées :



Numéros supprimés :

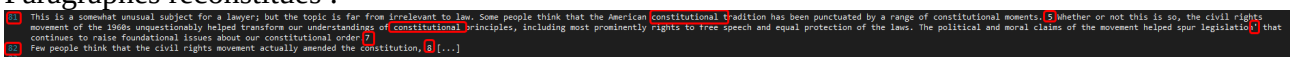


EXEMPLE 3 (RETOURS À LA LIGNE ET NUMÉROS DE RÉFÉRENCES)



Texte original :

Paragraphes reconstitués :



Numéros de références supprimés : (**Note :** on voit qu'il y a un 5, 7 et 8, mais pas de 6. Il faut chercher un peu, ici le 6 a été remplacé par un apostrophe après "legislation", donc il faut le supprimer aussi dans l'idéal)

EXEMPLE 4 (TRANSITION DE PAGE + NUMÉRO DE RÉFÉRENCE)
Texte original :

Texte nettoyé :