L1HN001 : Introduction aux humanités numériques
2e partie (resp. P. Amsili)
Devoir à la maison
à rendre au format PDF (iCampus)
avant le 13 décembre 2019 (minuit).

1. Récupérez l'extrait textuel qui vous a été assigné (voir dans le tableur) et faites en une copie dans votre document de travail (sous un traitement de textes).
2. **Segmentation** : appliquez strictement la règle de segmentation (1) sur votre extrait textuel (indiquez le début et la fin avec des crochets).

   (1) Une suite de caractères est une phrase si et seulement si elle débute par une majuscule et se termine par un point.

   Faites l'inventaire des difficultés rencontrées par la règle sur votre extrait, et proposez une nouvelle formulation aussi précise que possible qui marche parfaitement sur votre extrait. Montrer le bon fonctionnement de ces nouvelles règles sur une nouvelle copie de votre extrait.
3. **Tokenisation** : appliquez strictement la règle de tokenisation (2) sur votre extrait textuel (indiquez les séparations entre tokens avec une barre oblique (/).

   (2) Un token est toute suite de symboles différents de l'espace.

   Faites l'inventaire des difficultés rencontrées par la règle sur votre extrait, et proposez une nouvelle formulation aussi précise que possible qui marche parfaitement sur votre extrait. Montrer le bon fonctionnement de ces nouvelles règles sur une nouvelle copie de votre extrait.
4. **Catégorisation morphosyntaxique et lemmatisation** En présentant les 150 premiers tokens de votre extrait à raison d'un mot par ligne, et en utilisant le jeu d'étiquettes donné en cours [1], associez à chaque token une catégorie syntaxique (au moins une) et un lemme (au moins un). Discutez tous les cas problématiques, en appuyant votre discussion sur des exemples.

1. `http://ftb.linguist.univ-paris-diderot.fr/treebank.php?fichier=documentation`