Présentation du corpus de travail

Table des matières

[I. Les données brutes 1](#_Toc511002208)

[I.1 Obtention des données brutes auprès de HAL 1](#_Toc511002209)

[I.2 Chargement en mémoire et premières requêtes 1](#_Toc511002210)

[I.3 Constitution d’un lexique 1](#_Toc511002211)

[I.4 Suppression des titres en anglais 1](#_Toc511002212)

[I.5 Sérialisation 1](#_Toc511002213)

[I.6 Segmentation et étiquetage 1](#_Toc511002214)

[II. Le corpus travaillé 1](#_Toc511002215)

[II.1 Chargement 1](#_Toc511002216)

[II.2 Première requête 1](#_Toc511002217)

[Annexe A : informations sur le code 1](#_Toc511002218)

# I. Les données brutes

## I.1 Obtention des données brutes auprès de HAL

## I.2 Chargement en mémoire et premières requêtes

## I.3 Constitution d’un lexique

## I.4 Suppression des titres en anglais

## I.5 Sérialisation

## I.6 Segmentation et étiquetage

# II. Le corpus travaillé

## II.1 Format du fichier XML

<notices>

<notice>

<id>1712921</id>

<type>ART</type>

<date>2018</date>

<title>La logique de l'action de Michael Quante</title>

<words>

<word>La</word>

<word>logique</word>

<word>de</word>

<word>l'</word>

<word>action</word>

<word>de</word>

<word>Michael</word>

<word>Quante</word>

</words>

<pos\_tags>

<pos\_tag>DET</pos\_tag>

<pos\_tag>NOUN</pos\_tag>

<pos\_tag>ADP</pos\_tag>

<pos\_tag>DET</pos\_tag>

<pos\_tag>NOUN</pos\_tag>

<pos\_tag>ADP</pos\_tag>

<pos\_tag>PROPN</pos\_tag>

<pos\_tag>ADJ</pos\_tag>

</pos\_tags>

<authors>

<author>Alain Patrick Olivier</author>

</authors>

<domains>

<domain>0.shs</domain>

<domain>1.shs.phil</domain>

</domains>

</notice>

## II.2 Chargement

## II.3 Première requête

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **POSTAG1** | **POSTAG2** | **POSTAG3** | **POSTAG3** | **COUNT** | **% of total** | **% of titles with ":"** |
| \_\_EMPTY\_\_ | \_\_EMPTY\_\_ | \_\_EMPTY\_\_ | \_\_EMPTY\_\_ | 195846 | 68% |  |
| DET | NOUN | ADP | DET | 9028 | 3% | 10% |
| NOUN | ADP | DET | NOUN | 8348 | 3% | 9% |
| DET | NOUN | ADP | NOUN | 6106 | 2% | 7% |
| DET | NOUN | ADJ | ADP | 3522 | 1% | 4% |
| ADP | DET | NOUN | ADP | 2530 | 1% | 3% |
| NOUN | CONJ | NOUN | ADP | 2399 | 1% | 3% |
| NOUN | ADP | NOUN | ADP | 1830 | 1% | 2% |
| NOUN | ADJ | CONJ | NOUN | 1467 | 1% | 2% |
| DET | NOUN | ADP | PROPN | 1455 | 1% | 2% |
| DET | NOUN | ADJ | \_\_EMPTY\_\_ | 1359 | 0% | 2% |

|  |  |
| --- | --- |
| DET | déterminant |
| ADJ | adjectif |
| ADP | adposition (préposition ici) |
| CONJ | conjonction de coordination |
| NOUN | nom |
| PROPN | nom propre |
| \_\_EMPTY\_\_ | vide (pas d'élément) |

# Annexe A : informations sur le code