Projet Programmation Python,

Master 1 IdL – semestre 2 - année 2022-2023

Description

L'objectif de ce projet est de développer une série de fonctions Python permettant d'intégrer dans un même fichier CoNLL-U les résultats de différents analyseurs effectuant des chaînes de traitement variées (TOKENIZE, POS, LEMMA, DEPPARSE).

Pour ce faire il sera nécessaire de fournir en entrée un fichier tokenisé au format CoNLL-U ne contenant qu'une série de tokens numérotés. L'application de différents outils (Spacy, UDPipe, Stanza, HOPS) permettra alors de compléter différentes colonnes (LEMMA, POS, FEA, DEPREL). Les fichiers étant déjà tokenisés, la fusion des sorties ne nécessitera pas d'algorithme de synchronisation.

L'idée serait de pouvoir définir différents « processors » à la carte, à la manière de Stanza :

```
processors={
      "language":"fr",
      "tokenize" : {
               tool:"stanza",
               model:"standard"
      },
      "pos": {
               tool:"stanza",
               model:"partut"
      }
      "lemma": {
               "tool":"spacy",
               "model":"fr_core_news_sm",
      },
      "depparse": {
               "tool":"hops",
               "model":"UD_French-GSD-2.9-camembert"
      }
}
```

A partir de ces définitions, vous lancerez les différents outils pour lesquels vous aurez développé (ou récupéré) des fonctions *wrapper*, permettant d'en unifier les entrées et les sorties.
Vous fournirez par ailleurs une fonction de tokenisation maison capable de travailler sur des fichiers XML. Cette fonction aura pour tâche d'identifier les zones de texte à traiter (avec une formule XPATH) et de fournir en sortie une tokenisation CoNNL-U en plaçant sur la colonne 10 (MISC) les informations *SpaceAfter* (présence d'un espace après le token) et *Offset* (position du token dans le fichier source), afin de pouvoir lier chaque token à son document d'origine.

```
1    Par          _     _     _     _     _     _     Offset=996
2    exemple_     _     _     _     _     _     SpaceAfter=No|Offset=1000
3    ,      _     _     _     _     _     _     Offset=1007
```

Méthodologie (travail en binôme)

1. Phase d'imprégnation
2. Rédaction mini-cahier des charges
3. Développement et tests
    - Obligation de gestion et suivi du projet sous GitLab
    - Technologies : Python, Spacy, UDPipe, Stanza, HOPS
    - Serveur i3l
4. Démonstration à l'oral du travail réalisé à l'équipe enseignante

Rendus

- Le mini-cahier des charges
- Les différents scripts
- L'ensemble doit être déposé sur GitLab et décrit dans un fichier Readme.md