Qi WANG 21910034

Documents Structurés : Exercice 3 - TEI

Dans les premières deux séances du cours Documents Structurés, nous avons appris le langage XML avec ses règles, et la création de DTD (Document Type Definition) et TEI (Text Encoding Initiative) qui se trouvent dans des documents XML. Afin de tester et consolider nos connaissances, l'exercice 3 est proposé à la fin de la deuxième séance. Après plusieurs essais pénibles (souvent à cause de mon ignorance), enfin, je suis contente de pouvoir obtenir le fichier XML enrichi et valide. Dans des parties suivantes de ce document, je vais présenter mon travail par la chaîne de traitement, les difficultés et les solutions que j'ai pu résoudre ces problèmes.

• Chaîne de traitement

J'ai suivi principalement la chaîne de traitement proposée par Mme. Galleron, parce que cette proposition est explicite et efficace en tant que la guide de mon travail. Le traitement peut être divisé par 5 étapes en général:

- Créer le nouveau fichier de texte en éliminant les balises du fichier originel en XML. Il est à noter que la suppression du texte du header est nécessaire après l'enlèvement des balises, sinon ces métadonnées seront mélangées avec le contenu du texte dans le traitement suivant.
- Générer le fichier CoNLL-U (conllu) par UDPipe et le transformer en fichier csv.
- Nettoyer le fichier csv avec LibreOffice grâce aux expressions régulières (j'ai essayé OpenOffice mais il fonctionne moins bien sur ma machine, peut-être c'est à cause du système MacOS). L'idée principale est que nous devons supprimer des contenus moins utiles en gardant les informations importantes comme le numéro de phrase, les tokens, lemmas, pos, msd des phrases, etc. Ensuite, nous marquons ces informations avec les balises.
- Transformer le fichier csv en xml et nettoyer le fichier xml dans un éditeur de texte (j'ai utilisé Sublime Text). Dans cette étape-là, on se focalise à supprimer les espaces « blancs » comme les lignes vides, les retours à la ligne, les tabulations, etc.
- Travailler sous Oxygen. L'Oxygen examine s'il comporte des erreurs dans le fichier XML qu'on vient de produire, après on ajoute le header du fichier, et aussi les balises correspondantes aux chapitres dans le fichier originel.

• Difficultés et solutions

Le traitement fonctionne bien dans mon travail en pratique car il fournit des outils et des méthodes clairement indiqués. Néanmoins, j'ai quand même rencontré quelques problèmes.

- Quand je nettoyais le texte dans LibreOffice, je me suis rendue compte que l'annotation des ponctions doit être faite séparément par rapport à celle des tokens, parce que les ponctions sont entourées des balises différentes, soit <pc></pc> dans l'exemple attendu. Mais le problème est que les ponctuations sont aussi reconnues comme des tokens par UDPipe dont leurs pos sont « PUNCT ». Pour résoudre ce problème, j'ai d'abord supprimé les annotations « pourries » des ponctuations grâce aux expressions régulières : selon mes remarques précédentes, ces mauvaises annotations ont toujours des pos « PUNCT ». Ensuite, j'ai mis des balises <pc></pc> entre les ponctuations.

- À part l'élimination des espaces ou des retours à la ligne à l'étape 4 (travail dans l'éditeur de texte), beaucoup d'expressions régulières nécessitent d'être utilisées à cause de la transformation de csv à xml. Par exemple, nous pouvons observer que les symboles répétitifs ou inutiles (des virgules, des guillemets, etc) rendent le fichier xml invalide, d'après la capture d'écran ci-dessous. Donc j'ai remplacé les virgules par l'espace et aussi remplacé tous les deux guillemets par un seul guillemet.

• Conclusion

Par conséquent, ce travail est relativement long à faire mais je suis reconnaissante de résoudre tous les problèmes rencontrés. Non seulement j'ai bien compris la structure de TEI, mais aussi j'ai pu pratiquer des outils différents pour l'annotation des textes, et j'ai notamment familiarisé à la pratique des expressions régulières grâce à ce travail. De plus, ce travail m'a fait reconnaître l'importance de la patience et de la discrétion, qui seront aussi fortement demandés pour d'autres devoirs. Prenons un exemple, il faut toujours rassurer le résultat récent avant de passer à l'étape suivante.