04/02 reunion avec les profs

Notre travail : faire des annotations manuelles et automatiques (code) sur des Entités Nommées (NER) sur un corpus de poèmes déjà prêt. il faut choisir quels textes on va annoter (conseillé 1000 tokens par poème).

Nous allons utiliser la programmation pour l’analyse des données et nous allons utiliser certaines métriques (à décider) pour l'évaluation du modèle.

Question qu’on cherche à étudier : Est-ce que la nature du poème fait que le modèle fonctionne différemment ? (comparaison avec les romans)

Corpus : <https://github.com/These-SCAI2023/CORPUS_POEMES>

Observation : même s’il y a du bruit et des fautes d'orthographe sur le corpus, il ne faut pas les corriger !

Glossaire des humanités numériques : <https://digithum.huma-num.fr/ressources/glossaire/>

À faire pour le 28/02 :

1. Modèle large de spacy <https://spacy.io/usage/large-language-models>
2. Modèle Flauber t[ngface.co/docs/transformers/model\_doc/flaubert](https://huggingface.co/docs/transformers/model_doc/flaubert)
3. Modèle camembert

<https://almanach.inria.fr/software_and_resources/CamemBERT-fr.html>

1. Modèle Stanford <https://nlp.stanford.edu/software/CRF-NER.html>

Regarder les schéma d’annotation utilisés sur les 4 modèles.

Décider quel guide d’annotation nous correspond le mieux.

Créez vous un template latex de mémoire et créez le sur overleaf (https://www.overleaf.com/project/67b47266b94c349f24135b3d)

(Elles nous ont dit aussi de regarder les corpus qui ont déjà été utilisés avec ces modèles)

Mail de madame Caroline : Il faudrait que vous consultiez la documentation de spaCy pour la REN <https://spacy.io/usage/spacy-101> et que vous l'appliquiez au corpus mis à votre disposition.

Mail de Gaël lejeune : sélection des termes importants :

- Les **systèmes de Reconnaissance d’Entités Nommées** [...] semblent très sensibles à la variation des genres traités (presse, littérature…) [...](Lejeune 2023).  
- **modèles contextuels** (Luoma & Pysalo 2020).  
- On parle de **True casing**, Lita 2003)  
- l'objectif est d’**évaluer la performance** de systèmes état de l’art de REN  
- mesurer à quel point la **recherche d’un meilleur découpage en phrase** et d’une **casse plus fidèle** [...] contribuerait à améliorer les performances.

Appliquer un système de REN au corpus donné est donc la toute première étape.

18/02 réunion Lucine et Nina à la bibliothèque

Création du template mémoire sur overleaf :

Lien pour lecture :

<https://www.overleaf.com/read/rxgnbjskkrqn#134bcb>

Lien pour édition :

<https://www.overleaf.com/9614761656hdvgmjvzntjr#c2cb9a>

26/02 réunion à la bibliothèque

Familiariser avec le corpus :

Étudier chaque modèle et son schéma d’annotation :

Nina:

Modèle large de spacy <https://spacy.io/usage/large-language-models>

Modèle Flauber t[ngface.co/docs/transformers/model\_doc/flaubert](https://huggingface.co/docs/transformers/model_doc/flaubert)

Lucine:

Modèle camembert

<https://almanach.inria.fr/software_and_resources/CamemBERT-fr.html>

Modèle Stanford <https://nlp.stanford.edu/software/CRF-NER.html>

28/02 reunion avec les profs

* Articles sur CamemBERT :

<https://huggingface.co/Jean-Baptiste/camembert-ner>

<https://huggingface.co/docs/transformers/model_doc/camembert>

-> Ne pas besoin d'utiliser la tokenisation avec CamenBERT

* Utiliser google colab pour faire tourner spacy :

import spacy !python3 -m spacy download fr\_core\_news\_sm nlp = spacy.load('fr\_core\_news\_sm') doc = nlp("Je teste spacy à Paris") for word in doc.ents: print(word.text, word.label\_)

* Data sur google drive :

Pour importer les données stockées sur google drive il faut accepter les conditions d'utilisation à chaque fois

from google.colab import drive drive.mount('/content/drive')

<https://stackoverflow.com/questions/48376580/how-to-read-data-in-google-colab-from-my-google-drive>

À faire pour le 14/03 :

Exécution des modèles (surtout Spacy) une fois

Étapes pour le code : Chemin vers fichier, lire fichier, chargement et exécution du modèle SpaCy, extraction des entités nommées et stocker les données (sorties sous forme de dictionnaire) quelque part (on peut les sauvegarder en format Json).