Modèles de langage basés sur l'architecture *Transformer* pour le français

Encadrant : francois.role@parisdescartes.fr

Contexte
Des modèles de langage du français ont tout récemment été publiés.

CamemBERT: a Tasty French Language Model
https://arxiv.org/abs/1911.03894
https://camembert-model.fr
https://huggingface.co/transformers/model_doc/camembert.html

FlauBERT
https://arxiv.org/abs/1912.05372
https://github.com/getalp/Flaubert

C'est positif car même des outils avancés comme Spacy sont notoirement encore peu adaptés à traiter certaines langues dont le français. Il reste à évaluer ce que ces modèles peuvent vraiment apporter.

Travail à réaliser

1. Reproduire avec l'implémentation
https://huggingface.co/transformers/model_doc/camembert.html

les manipulations décrites dans https://github.com/pytorch/fairseq/tree/master/examples/camembert

2. Tester la capacité des deux modèles cités plus haut à s'adapter à des corpus spécialisés en effectuant un *fine-tuning* sur un corpus de données spécialisées.

Outils

- CamemBERT : implémentation Fairseq et Huggingface (transformers)
- FlauBERT : Huggingface (transformers)

Rendu
Un code clair et commenté dans un notebook. Un rapport identifiant et décrivant précisément les points techniques difficiles de la tâche 1 et décrivant en détail le plan d'expérience pour la tâche 2.

Critères d'évaluation
Qualité du code. Qualité de la rédaction du rapport. Qualité de l'expérience de *fine-tuning*.