Modalités d'évaluation

Gustave Cortal

Un projet en équipe portant sur **un** (et non plus au moins deux) jeu de données que vous aurez choisi et qui aura été validé par moi-même. Voici les différentes étapes à suivre :

- Présentation du jeu de données (cf. datasheet cours 1) [2pts]
- Pré-traitement du jeu de données [4pts] :
  - appliquer la tokénisation à base d'expressions régulières (votre propre tokeniser regex ou utiliser NLTK) et la tokénisation byte-pair encoding (avec sentencepiece, tiktoken ou huggingface), cf. cours 1 et 2.
  - appliquer des méthodes de normalisation du texte comme la suppression des stop words, la lemmatisation et le fait de tout mettre en minuscule, cf. cours 1 et 2.
- Statistiques descriptives sur vos données [2pts] : nombre de documents, phrases, tokens, classes à prédire, les tokens les plus fréquents, etc.

- Entrainement de plusieurs modèles prédictifs sur votre jeu de données avec vos propres implémentations ou en utilisant des bibliothèques comme NLTK et scikit-learn [10pts] :
  - Entrainement de : n-gram (cours 2), bayésien naïf (cours 3), régression logistique (cours 4), tf-idf et word2vec (cours 5), réseaux de neurones feedforwards (cours 6), réseaux de neurones récurrents (derniers cours, optionnel, pour des points bonus), transformer (derniers cours, optionnel, pour des points bonus).
  - Évaluer les performances de vos modèles entrainés et comparer les avec des métriques comme la perplexité, le recall, la precision, le f1-score, etc. (cours 3).
  - Varier plusieurs configurations d'entrainement pour évaluer l'impact de certains choix sur les performances. Par exemple, varier la façon de pré-traiter les données, varier les hyperparamètres de vos modèles, etc.
- Limitations de vos approches, difficultés rencontrées et pistes d'améliorations [2pts]

Points bonus sur les manières créatives d'aborder les étapes : interprétation des modèles, optimisation des hyperparamètres, évaluation avec validation croisée, transfert de connaissance en croisant plusieurs jeux de données, etc.

Les groupes devront rendre un rapport écrit sur LateX qui rend compte des différentes étapes. Deadline pour la rendu du projet écrit le **20 mai**.

Les groupes devront aussi faire une présentation orale de 10 min durant le dernier cours, le **27 mai**.