

# Modalités d'évaluation

Gustave Cortal

# Modalités d'évaluation

Un projet en équipe (3 à 6 personnes) portant sur un jeu de données que vous aurez choisi et qui aura été validé par moi-même.

Les groupes rendent un **rapport écrit**. Le rapport ne doit **pas dépasser 10 pages**. Deadline pour l'envoi du rapport le **12/05**. Envoyer le pdf du rapport par mail à [gustavecortal@gmail.com](mailto:gustavecortal@gmail.com) avec comme titre *NLP - Groupe [numéro inscrit sur l'excel]*. Ils effectuent une **présentation orale** lors du dernier cours, le **19/05**.

# Modalités d'évaluation

Dans le rapport, vous devez rendre compte de manière concise de plusieurs étapes pour l'élaboration d'un projet NLP complet :

- ▶ Présentation du jeu de données *2.5 pts*
- ▶ Pré-traitement du jeu de données (normalisation et tokenisation) et analyse exploratoire (statistiques descriptives sur les documents, classes, tokens, etc.) *2.5 pts*
- ▶ Benchmark sur les modèles de classification (bayésien naïf, régression logistique, tf-idf, word2vec, réseaux de neurones feedforwards, réseaux de neurones récurrents, transformer) *5 pts*
- ▶ Benchmark sur les modèles de génération de texte (n-gram, tf-idf et word2vec, réseaux de neurones feedforwards, réseaux de neurones récurrents, transformer) *5 pts*

J'évalue aussi la clarté (présentation des objectifs, mise en page, agencement des idées) du rapport, votre capacité à aller à l'essentiel (ne pas donner d'informations inutiles à la compréhension, par exemple, pas besoin de réexpliquer ce qu'est une régression logistique). J'évalue aussi comment vous avez identifié les limites de vos approches, les difficultés rencontrées, les pistes d'amélioration. *5 pts*

# Modalités d'évaluation

Vous devez avoir exploré en profondeur trois de ces approches (liste non-exhaustives) :

Augmenter des données (au moins trois méthodes différentes, sans et/ou avec des modèles de langues) et évaluer leurs impacts sur les performances, créer votre jeu de données (par exemple, avec du scrapping), combiner plusieurs jeux de données et évaluer leurs impacts sur les performances, interpréter les modèles avec au moins deux méthodes différentes, effectuer de l'apprentissage non supervisé avec au moins deux méthodes différentes (par exemple, réduction de dimension puis clustering et/ou topic modeling), combiner plusieurs modèles ensemble avec au moins deux méthodes différentes (bagging, stacking, modèles hiérarchiques), créer une application (par exemple, développement d'une interface graphique) *7.5 pts*

Le rapport et la présentation orale doivent rendre compte de ces approches et les indiquer clairement.

La présentation orale doit présenter de manière concise votre rapport écrit pendant 10 minutes maximum. *2.5 pts*

# Conseils

Les benchmarks peuvent être concis, représentés dans un tableau qui compare les performances des modèles. Les résultats des benchmarks doivent être interprétés. Les choix des modèles et des configurations doivent être justifiés. Il faut varier plusieurs configurations d'entraînement comme les manières de pré-traiter les données et les hyperparamètres des modèles. Le but est de trouver les meilleures performances pour chaque modèle.

Il est possible d'entraîner d'autres modèles que nous n'avons pas vus en cours comme XGBoost, Random Forest, etc.

Vous pouvez utiliser toutes les bibliothèques python disponibles. Si j'ai des doutes sur vos résultats, je demanderai à lire votre code.