



OBJECTIFS

Objectifs pédagogiques

Se familiariser avec les différentes étapes d'un projet de traitement du langage naturel (preprocessing, feature representation, choix/création du modèle, évaluation), sur un problème classique : l'analyse de sentiments. On implémentera deux approches différentes, une classique, et une de deep learning sur un jeu de données de commentaires de films, pour essayer de prédire si ces commentaires sont positifs ou négatifs.

Démarche pédagogique (projet, ressources ...)

- Durée du projet : 3 jours
- Travail en autonomie, mais échangez autant que possible entre vous !
- Produire vos propres scripts et mémos individuels pour terminer le projet

Compétences :

- Maitriser les étapes de preprocessing pour le traitement du langage naturel
- Implémenter un classifieur utilisant une représentation en bag-of-words
- Implémenter un modèle de deep learning (LSTM) pour faire de la classification de texte
- Réutiliser des word embeddings pour initialiser un modèle de deep learning

MODALITÉS

Durée

3 jours soit 21 heures au total.

Lancement le 1/12/25 et clotûre le 3/12/25.

Formateur(s)

Théo Trouillon

Livrables

- Le notebook rempli

Ressources

- Bag-Of-Word and TF-IDF:

<https://www.analyticsvidhya.com/blog/2020/02/quick-introduction-bag-of-words-bow-tf-idf/>

- Recurrent Neural Networks (RNNs):

<https://medium.com/towards-data-science/illustrated-guide-to-recurrent-neural-networks-79e5eb8049c9>

- Long Short Term Memory networks (LSTMs):

<https://medium.com/towards-data-science/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21>

<http://colah.github.io/posts/2015-08-Understanding-LSTMs/>

- Word embeddings:

<http://jalammar.github.io/illustrated-word2vec/>