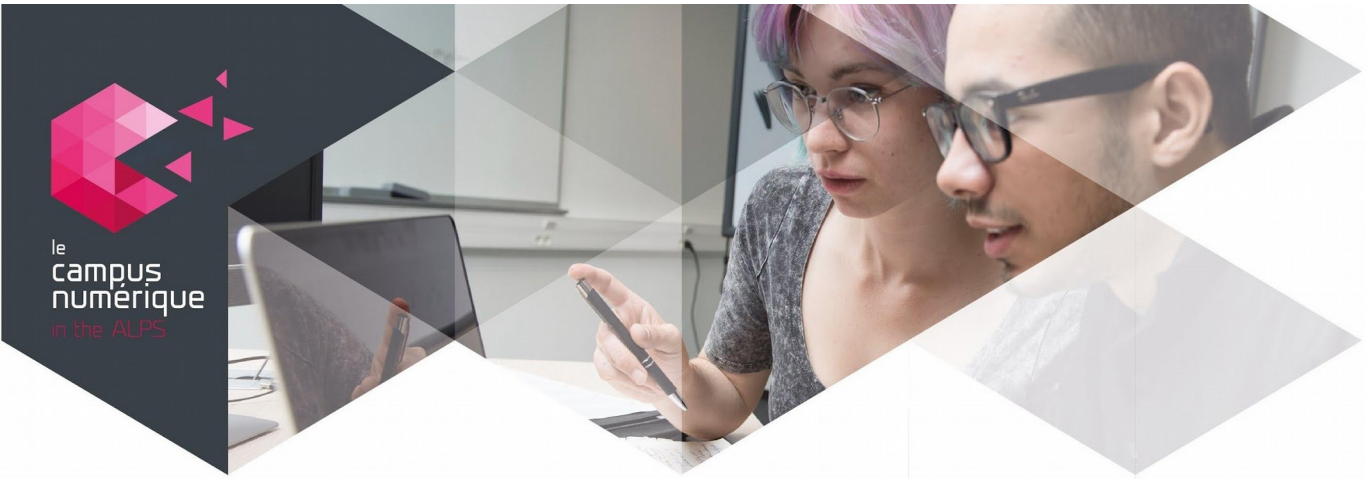

MACHINE LEARNING
Natural Language Processing

DATA+ 2023 #2

Référent module : Théo Trouillon

Objectifs pédagogiques

A l'issue de ce module, vous serez capable de :

- Effectuer les étapes de pré-processing nécessaire au nettoyage d'un texte pour du traitement du langage naturel (NLP)
- Utiliser des représentation de texte bag-of-word pour encoder du texte
- Implémenter un réseau de neurone récurrent (LSTM) pour faire de la classification de texte
- Réutiliser des embeddings de mots pré-entraîner pour accélérer l'entraînement de vos modèles de deep learning

Pré-requis

- Programmation en Python
- Bases d'algèbre linéaire
- Bases de machine learning (classification)
- Bases de deep learning

Projet : Natural Language Processing (3 jours)

Modalités

- Travail en autonomie
- Production individuelle

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

Consignes

- Ouvrir et compléter le notebook

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://towardsdatascience.com/illustrated-guide-to-recurrent-neural-networks-79e5eb8049c9
- Long Short Term Memory networks (LSTMs): https://towardsdatascience.com/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/

Livrables

- ❑ Le notebook rempli