3MD4150 - Traitement du langage naturel

Responsables : **Joël LEGRAND**

Langues d'enseignement : **FRANCAIS**

Campus où le cours est proposé : **CAMPUS DE METZ**

Nombre d'heures d'études élèves (HEE) : **40**

Nombre d'heures présentielles d'enseignement (HPE) : **29**

Année académique : **2024-2025**

Niveau avancé : **non**

Présentation, objectifs généraux du cours :

Le traitement automatique du langage (TAL) est une discipline, à la croisée de l'apprentissage automatique et de la linguistique, permettant d'exploiter automatiquement des données textuelles en langage naturel à l'aide d'outils informatiques.
Cet enseignement vise à introduire les concepts linguistiques, les méthodes et les outils permettant de manipuler et d'exploiter de grandes quantités de données textuelles.

Période(s) du cours (n° de séquence ou hors séquence) :

SM11

Prérequis :

- Maîtriser les concepts de base de l'apprentissage automatique
- Avoir une experience d'utilisation de librairie d'apprentissage profond (Tensorflow, pytorch, torch, ...)

Plan détaillé du cours (contenu) :

Cet enseignement introduit les principales théories linguistiques permettant de modéliser le langage naturel (ex: grammaires formelles, grammaires de dépendances, ...).
Il présente les différents outils de traitement automatique de langues (TAL) disponibles ainsi que modèles statistiques à la base de ceux-ci.
L'accent sera notamment porté sur les méthodes d'apprentissage profond qui constituent l'état de l'art pour la plupart des tâches de TAL.

Déroulement, organisation du cours :

Chaque séance comprendra une partie de cours magistral (CM) au cours duquel de nouvelles notions seront introduites, suivi d'une séance de travaux pratiques (TP) sur machine. Les TP seront des applications directes des notions vues en CM. L'ensemble du matériel pédagogique (support de CM et de TP) sera fourni aux étudiants.

Organisation de l'évaluation :

Deux notes seront prises en compte pour l’évaluation de cet enseignement.

Evaluation 1:
Type d’examen: Examen sur machine
Acquis d'apprentissage évaluées: Utilisation des notions linguistiques et des outils informatiques introduits en cours.
Modalité: L’examen aura lieu en salle machine et comprendra une partie théorique sur les modèles statistiques à la base des outils de TAL. Il sera suivi d’une partie pratique de mise en application desoutils et concepts linguistiques vus en cours, appliqués sur un problème réel.
Pourcentage: 50%

Evaluation 2: Projet
Type d’examen: Projet à rendre
Acquis d'apprentissage évaluées: Utilisation des notions linguistiques et des outils informatiques introduits en cours. Faire preuve d’autonomie et de créativité face à un problème concret.
Modalité et retour: Le sujet ainsi que le barème de notation seront communiqués en milieu de module. Il s’agira d’exploiter des données issues de la plate-forme Kaggle fournissant des challenges en science des données, sur la base de problématiques industrielles réelles. Le projet sera à rendre en fin de module. Un retour sur le travail fourni sera rendu en même temps que la note.
Pourcentage: 50%

Moyens :

Les cours magistraux (CM) seront assurés pas Joël Legrand et les scéances de travaux pratiques (TP) par Joël Legrand et Jérémy Fix.
Les notions théoriques seront introduites en cours magistral (CM) puis mise en application lors de scéances de TP sur machine.
Les TP se feront en majorité en python; les TP d'apprentissage profond se feront à l'aide de la librairie PyTorch (https://pytorch.org/).

Acquis d'apprentissage visés dans le cours :

- Se familiariser avec les bases théoriques permettant de conceptualiser et de modéliser les phénomènes linguistiques.
- Maîtriser les outils essentiels du TAL (lemmatiseur, analyseur syntaxiques, etc.).
- Acquérir une autonomie pour le traitement automatique de contenus textuels.

Description des compétences acquises à l'issue du cours :

C1.4, Marker 2 : Design, detail and corroborate a whole or part of a complex system.
C3.6, Marker 1 : Evaluate the efficiency, feasibility and strength of the proposed solutions.
C8.1, Marker 3 : Work collaboratively in a team.

Bibliographie :

- Installation de la librairie d’apprentissage profond PyTorch: https://pytorch.org/
- Livre de référence sur l’apprentissage profond (en anglais): https://www.deeplearningbook.org/
- Les outils d’analyse linguistique du groupe de TAL de Stanford: https://nlp.stanford.edu/software/
- La documentation du package NLTK pour Python: https://www.nltk.org/