Module : Text Mining and NLP				Code	
				ING-5-SDIA-S9-P3	
Période	**S9-P1**	*Volume horaire*	**42h**	*ECTS*	**4**

Responsable	**Sonia Gharsalli**	*email*	Sonia.gharsalli@tek-up.tn
Equipe pédagogique	Sonia Gharsalli		

1. Objectifs de Module *(Savoirs, aptitudes et compétences)*

Ce module porte sur le traitement de Text et les méthodes de machine learning et de deep learning favorisant l'extraction de connaissance à partir de texte.

Acquis d'apprentissage :

A la fin de cet enseignement, l'élève sera capable de :

- Maîtriser les notions de base pour le traitement de l'information textuelle (**C1.2**)
- Simuler et tester des méthodes de classification de texte et d'extraction des thématiques d'un texte (**C1.3**)
- Concevoir des solutions basées sur le traitement de l'information textuelle (**C1.1**)
- Communiquer des produits complexes …… (**C3.3**)

Compétences
C1.2 Concevoir des produits (modèles) pour résoudre des problèmes pour les entreprises : extraction automatique des entités (mots clés dans le domaine financier…)
C1.3 Résoudre des problèmes complexes (ex : rechercher les ressemblances entre des décret de loi…)
C3.2 Développer un sens de critique des solutions existantes et capaciter de développement de solution innovatrice.

2. Pré-requis*(autres UE et compétences indispensables pour suivre l'UE concernée)*

- Programmation Python
- Notions de machine learning
- Expérience sur les framework deep learning (tensorflow, pytorch)

3. Répartition d'Horaire de Module

Intitulé de l'élément d'enseignement	*Total*	*Cours*	*TD*	*Atelier*	*PR*

Module : Text Mining and NLP	42	24		18	

4. Méthodes pédagogiques et moyens spécifiques au Module
(pédagogie d'enseignement, ouvrages de références, outils matériels et logiciels)

- Supports de Cours
- Projecteur et Tableau
- Travaux dirigés
- Travaux pratiques

Bibliographie

Titre	Auteur(s)	Edition
Text Mining : Applications and Theory	Michael W. Berry, Jacob Kogan	John Wiley & Sons/2010
Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network	Alex Sherstinsky	Elsevier "Physica D: Nonlinear Phenomena" journal, Volume 404, March 2020: Special Issue on Machine Learning and Dynamical Systems
Bert: Pre-training of deep bidirectional transformers for language understanding	Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. .	Article 2018
Mastering NLP from Foundations to LLMs Apply Advanced Rule-based Techniques to LLMs and Solve Real-world Business Problems Using Python	Lior Gazit, Meysam Ghaffari	ISBN :9781804616383, Publication :26 avril 2024

5. Contenu *(Descriptifs et plans des cours / Déroulement / Détail de l'évaluation de l'activité pratique)*	Durée allouée	
Séance 1 ● Introduction au Text Mining et applications ● Matrice document Termes : (BOW, pondération, mesure de similarité, N-Gramme)	Cours	3H
Séance 2 ● Catégorisation de texte (métriques d'évaluation des performance, Réduction de dimension, méthodes de classification)	Cours	3H
Séance 3	Atelier	3H

• Extraction et Nettoyage d'un corpus de texte à partir de la bibliothèque NLTK et utilisation des méthodes de classification		
Séance 3 • Topic Model (introduction de la notion, algorithme Latent semantic indexing (LSI), Analyse Factorielle des correspondances (AFC), Latent Dirichlet Allocation (LDA))	Cours	3H
Séance 4 • Recherche des thématiques dans un texte en utilisation la méthode LDA	Atelier	3H
Séance 5 • Word Embedding : la représentation Naïve du prolongement des termes, exemple introductif au prolongement, Continuous BOW, Skip Gram)	Cours	3H
Séance 6 • Recherche de similarité entre des termes du vocabulaire en utilisation l'algorithme word2vec • Word Embedding en utilisant Spacy	Atelier	3H
Séance 7 • Modèles séquentiels : RNN, LSTM,GRU	Cours	3H
Séance 8 • Classification des sentiments en utilisant les modèles séquentiels (tensorflow)	Atelier	3H
Séance 9 • Séquence to sequence models • Traduction de texte	Cours + Atelier	3H
Séance 10 • Transformers architecture, notion d'attention • Bert, T5, GPT2	Cours	3H
Séance 11 • Bot : Génération de texte en utilisant gpt2	Atelier	3H
Séance 12 • Introduction aux LLMs Vue d'ensemble des fonctionnalités avancées des LLM - Cas d'utilisation des capacités avancées des LLM	Cours + Atelier	3H
Séance 13 : • Fonctionnalités avancées des LLM • AI Agent • Retrieval augmented generation fonctionnement et exemples	Cours	3H

Séance 14 • Retrieval augmented generation • Langchain, deeplake,…	Atelier	3H

6. Mode d'évaluation de Module*(nombre, types et pondération des contrôles)*

Eléments d'enseignement	Coeff	DS	EX	TP	PR
Module - Text Mining and NLP	2	40%	60%		

Pour valider le module, les étudiants passeront un examen dont le coefficient est de 60%, un DS dont le coefficient est de 40%.

La durée de tous les examens (Examen, DS…) est de 1h30.

Le DS est planifié 7 semaines après le début du module.

Quant à l'examen, il est planifié après l'écoulement des 14 semaines et portera sur toutes les thématiques enseignées tout au long des 42 heures.

Le module est validé si l'étudiant obtient une moyenne supérieure ou égal à 10 sur 20.