INTITULE DU COURS : Data Science

CODE : IA416, GL24

COEFFICIENT: IA 1, GL 2

CREDITS : IA 2, GL 3

VOLUME HORAIRE HEBDOMADAIRE : 1h30 Cours, 3h00 TP

DUREE SEMESTRIELLE TOTALE DU COURS : 15 semaines (22h30 Cours, 22h30 TP)

FILIERE/SPECIALITE : Informatique-M1 (GL/IA)

LANGUE DU COURS : Français

CHARGE DE COURS : Bambrik Ilyas

EVALUATION : M1IA CC 40% , Examen 60% ; M1GL CC 34% , Examen 66%

CONTACT

Affiliation : Enseignant chercheur, Département d'informatique, Laboratoire LRIT,

Mail : ilyas.off.script@gmail.com

Domaines d'expertise : Algorithmique et Structure de donnée, Développement Web

Disponibilité : Jeudi et Samedi 17h00 – 18h00 en ligne sur Teams

PRESENTATION DU COURS

L'objectif général de ce cours est d'introduire l'étudiant au domaine de la science des données. En parallèle, l'étudiant sera initié à la programmation réseau avec Python.

OBJECTIFS D'APPRENTISSAGE

A l'issue de ce cours l'apprenant doit être capable de:

- Appliquer les fonctionnalités pandas pour transformer les données.
- Distinguer entre les différents types de graphiques.
- Utiliser seaborn pour visualiser les données tabulaires.
- Manipuler les données géographiques.

DESCRIPTIF ET STRUCTURE

- **Chapitre I Introduction à Pandas:**

Bref introduction aux objets Series et Dataframe.

- **Chapitre II Transformation des données:**

PRE-REQUIS

Les connaissances préalables que l'étudiant doit avoir afin de suivre ce cours :

- Comprendre les notions générales des Statistiques et Probabilités **(Cours L3)**;

RESSOURCES

Les ressources suivantes sont recommandées comme complément du contenu du cours:

[1] Datacamp: https://www.datacamp.com/, Cours gratuits

[2] Kaggle : https://www.kaggle.com/learn/, Cours gratuits

[3] Freecodecamp: https://www.freecodecamp.org/learn/data-analysis-with-python/#data-analysis-with-python-course

[4] HackerRank.com : https://www.hackerrank.com/domains/python?filters%5Bstatus%5D%5B%5D=unsolved&badge_type=python

ORGANISATION COURS

Le cours aura lieu chaque Mardi à 11h30, faculté des Sciences – Salle N101. Le déroulement du cours, TD et TP sera comme suit :

- Chaque séance de cours commence par 10 minutes de rappel.
- L'entré en cours n'est pas permise pour un retard supérieur à 10 minutes.
- Deux semaines pour la réalisation de chaque série TP. Après le début d'un nouveau TP, la correction type est fournie dans la vidéo explicative.
- A la fin d'une séance TP, chaque étudiant est responsable d'éteindre son PC.
- Deux ***teste TP sur feuille*** sont programmés durant le semestre.
- Chaque séance de TP commence par 15 minutes réservées aux questions posées par les étudiants.

CONSIGNES POUR LES EXERCICES OU TRAVAUX, INDIVIDUELS OU DE GROUPE

- Au début de chaque séance, la progression dans la série TP est notée.
- Le travaille en équipe est permis. Cependant, chaque étudiant est évalué individuellement par des questions concernant la solution proposée.
- Le délai de soumission d'un devoir doit être respecté. Tout retard dans la remise de devoir sera sanctionné **(-4 de la note finale pour chaque semaine de retard)**.

EVALUATION

Ce chapitre couvrira différentes transformations que nous pouvons appliquer à nos données ainsi que les fonctions de sommaire.

- **Chapitre III Data Cleaning:**

Exploite les méthodes de détection des inconsistances dans les données et des entrées manquantes.

- **Chapitre IV Visualisation des données:**

Dans ce chapitre explore seaborn et la visualisation de données.

- **Chapitre V Web Scraping:**

Ce chapitre couvrira la récolte de données avec requests et BeautifulSoup.

- **Chapitre VI Données Géospaciales:**

Ce chapitre introduit geopandas qui est un module développé spécialement pour la visualisation et manipulation des données géospatiales.

- **Chapitre VII Big Data:**

Ce chapitre introduit le paradigme Mapreduce et le traitement parallèle du Big Data avec pySpark.

MATERIEL DE COURS

LOGICIELS :

a) Anaconda :

https://repo.anaconda.com/archive/Anaconda3-2019.03-Windows-x86_64.exe

MATERIEL :

a) Les étudiants qui ne disposent pas d'un PC portable doivent se procurer d'un FlashDisk ou d'un périphérique de stockage afin sauvegarder leurs travaux à la fin de séance.

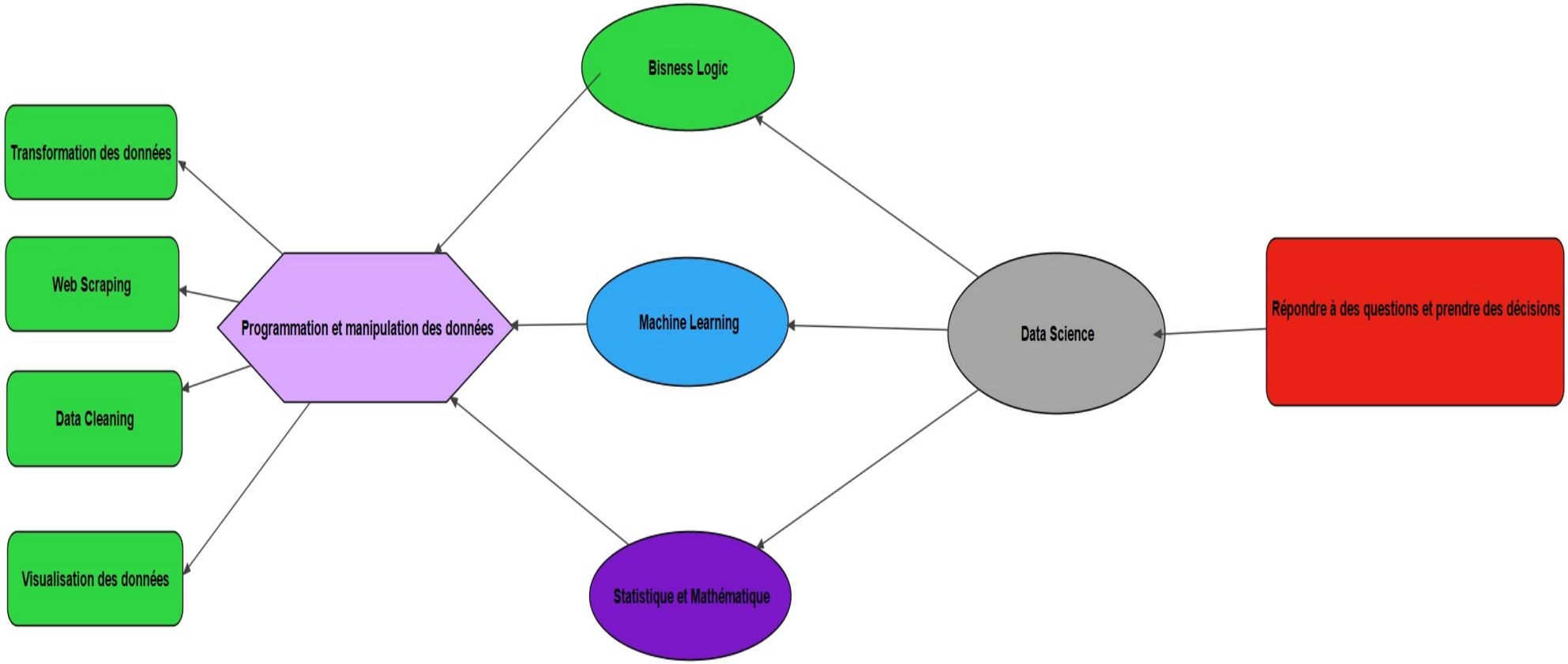
b) Il est préférable de préparer le devoir TP sur PC portable ou bien de l'exporter sur mémoire de stockage externe afin de l'exécuter sur un PC du laboratoire lors de la consultation. Cependant, c'est permis d'apporter le devoir sur papier pour la consultation.

MODULES CONNEXES :

Les connaissances acquises dans ce module seront très utiles dans les modules suivants :

- Analyse de données
- Intelligence artificielle

- ***La consultation des devoirs TP se concentre principalement sur l'analyse / synthèse. L'affichage de du résultat est secondaire*** ;
- La note finale de ce module est réparties sur : a) note TP (coefficient 1), b) note contrôle (coefficient 1), c) note d'examen (coefficient 3) ;
- Chaque étudiant doit avoir ça pièce d'identité lors de l'examen /CC;
- **Il est interdit de rependre dans une copie d'examen / teste avec un crayon ;**
- L'examen finale est d'une durée de **1h15,** composé de questions de cours seulement;
- Deux testes TP **écrits** de 45 minutes sont programmés au cours du semestre s ;
- La note TP finale est composée de la moyenne des deux testes TP sur 14 + la note des travaux TP et assiduité sur 6.
- **Le copiage des TP ou dans l'examen est pénalisé pour toutes les parties impliquées ;**


Bisness Logic
Transformation des données
Web Scraping
Programmation et manipulation des données
Machine Learning
Data Science
Répondre à des questions et prendre des décisions
Data Cleaning
Visualisation des données
Statistique et Mathématique