Web Mining

Course title – Intitulé du cours	Web Mining
Level / Semester – Niveau /semestre	M2/S2
School – Composante	Ecole d'Economie de Toulouse
Teacher – Enseignant responsable	PITARCH Yoann
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Other teacher(s) – Autre(s) enseignant(s)	
Lecture Hours – Volume Horaire CM	21
TA Hours – Volume horaire TD	
TP Hours – Volume horaire TP	
Course Language – Langue du cours	English
TA and/or TP Language – Langue des TD et/ou TP	English

Teaching staff contacts – Coordonnées de l'équipe pédagogique :

Yoann Pitarch – pitarch@irit.fr

Course Objectives – Objectifs du cours :

Web Mining can be seen as the adaptation of data mining and machine learning to specific web data. In reality, Web Mining is commonly divided into 3 distinct categories depending on the data being manipulated.

1. Web Content Mining is concerned with the analysis of the content that can be found on the Web. In this course we will focus only on textual content and will address issues such as topic extraction or sentiment detection.
2. Then, Web Structure Mining is interested in the analysis of the links that exist between the entities that populate the Web. These entities can be users of social networks linked by friendship links or web pages linked together by clickable links. We will then be interested in problems such as the discovery of important entities in this network or the extraction of user communities.
3. Finally, Web Usage Mining focuses on the analysis of traces left by users on the Web to for example recommend products to buy or build user profiles.

In this course, we will address these 3 sub-domains and introduce some techniques to address the issues mentioned.

Le Web Mining peut être vu comme l'adaptation de la fouille de données et de l'apprentissage automatique aux données particulières du Web. En réalité, le Web Mining est communément divisé en 3 catégories bien distinctes en fonction des données manipulées.

1. Le Web Content Mining s'intéresse à l'analyse du contenu que l'on peut trouver sur le Web. Dans ce cours nous nous intéresserons seulement au contenu textuel et aborderons des problématiques telles que l'extraction de sujets ou la détection de sentiments.
2. Ensuite, le Web Structure Mining s'intéresse quant à lui à l'analyse des liens qu'il existe entre les entités qui peuplent le Web. Ces entités peuvent être des utilisateurs de réseaux sociaux liés par des liens d'amitié ou des pages internet liée entre elles par des liens HTML. Nous intéresserons alors à des problématiques telles que la découverte d'entités importantes dans ce réseau ou l'extraction de communautés d'utilisateurs.
3. Enfin, le Web Usage Mining se focalise sur l'analyse des traces laissées par les utilisateurs sur le Web pour par exemple recommander des produits à acheter ou construire des profils utilisateurs.

Dans ce cours, nous aborderons ces 3 sous-domaines et introduiront quelques techniques pour répondre aux problématiques mentionnées.

Prerequisites – Pré requis :

- Machine Learning (basics)
- Graph theory
- Python for Data Science

Practical information about the sessions – Modalités pratiques de gestion du cours :

Most sessions will begin with a presentation of the session topic and end with a practical application of the concepts introduced.

La plupart des sessions de cours débuteront par un exposé introduisant des concepts associés au sujet de la séance et se poursuivront par une mise en pratique de ces concepts.

Grading system – Modalités d'évaluation :

A project (Kaggle competition) will be used to evaluate the students.`

Un projet (compétition Kaggle) servira d'évaluation à ce projet.

Bibliography/references – Bibliographie/références :

- Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data by Bing Liu
- Mining the Social Web by Matthew A Russell and Mikhail Klassen
- Python Data Science Handbook by Jake VanderPlas

Session planning – Planification des séances

1. Introduction and Prerequisites

2. Web Content Mining - 1
3. Web Content Mining - 2
4. Web Structure Mining - 1
5. Web. Structure Mining - 2
6. Project Q&A
7. Web Usage Mining

Distance learning – *Enseignement à distance :*

Distance learning can be provided when necessary by implementing :

- *Interactive virtual classrooms*
- *Remote (online) tutorials (classes)*
- *Chatrooms*

En cas de nécessité, un enseignement à distance sera assuré en mobilisant:

- *Classe en ligne interactive*
- *TP/TD à distance*
- *Forum...*