**Description du module, disponible en: FR**

# *Recherche d'informations et Data Mining*

## Informations générales

**Nombre de crédits ECTS**

3

**Code du module**

TSM_InfData

**Valable pour l'année académique**

2020-2021 DRAFT

**Dernière modification**

2020-01-31

**Nom du/de la responsable de module**

Laura Raileanu (HES-SO, Laura.Raileanu@heig-vd.ch)

**Explication des définitions de langue par lieu :**

- Les cours se dérouleront dans la langue définie ci-dessous par lieu/exécution.
- Les documents sont disponibles dans les langues définies ci-dessous. Pour le multilinguisme, voir la répartition en pourcentage (100% = documents complets)
- L'examen est disponible à 100% dans chaque langue sélectionnée pour chaque lieu/exécution.

| | Berne | Lausanne | | | Lugano | Zurich | | |
|---|---|---|---|---|---|---|---|---|
| **Leçons** | | | **X** F 100% | | | | | |
| **Documentation** | | **X** E 100% | **X** F 20% | **X** E 80% | | | | |
| **Examen** | | | **X** F 100% | | | | | |

**Catégorie de module**

TSM approfondissement technico-scientifique

**Leçons**

2 leçons et 1 leçon de pratique par semaine

## Compétences préalables

**Connaissances préalables, compétences initiales**

- Connaissances dans le domaine de bases de données relationnelles
- Connaissances de base en statistiques
- Bonnes bases de programmation orientée objets (Java)

## Brève description du contenu et des objectifs

-

## Objectifs, contenus, méthodes

### Objectifs d'apprentissage, compétences à acquérir

- Le cours fournit une introduction au domaine de la recherche d'informations et au domaine multidisciplinaire de data mining.
- Les étudiants connaissent l'architecture d'un système de recherche d'information.
- Ils connaissent les modèles de RI (booléen et vectoriel), ainsi que l'utilisation de ces modèles pour déterminer le poids des termes d'indexation et calculer la correspondance entre les documents et les requêtes.
- Ils comprennent les différentes mesures de l'évaluation d'un système de recherche d'information et sont capables d'appliquer les algorithmes de comparaison et d'interpréter leurs résultats.
- Ils connaissent l'utilisation de la librairie Apache Lucene pour l'indexation et la recherche d'information selon le modèle booléen et vectoriel
- Ils connaissent les techniques de détection des documents similaires en utilisant les algorithmes de type « Localitiy Sensitive Hashing »
- Les étudiants comprennent l'utilisation des technologies modernes de bases de données pour le traitement et la gestion de grandes collections de données.
- Les étudiants reçoivent une introduction au domaine de bases de données multidimensionnelles, aux modèles d'entreposage de données, aux techniques OLAP. Ils connaissent de nouvelles structures de données (types de données) alternatives aux systèmes de gestion de bases de données relationnels (SGBDR) (non relationnelles notamment) et sont capables de déterminer quels types de données et quel système de base de données sont appropriés en fonction du contexte et du genre de données disponibles.
- Ils connaissent les techniques de pré-traitement des données (le concept de qualité des données et les méthodes de nettoyage des données, d'intégration des données, de réduction des données, de transformation des données et de discrétisation des données).
- Ils connaissent les principales tâches de data mining et les méthodes principales associées : analyse descriptive de données, analyse du panier d'achats (règles d'association), classification (arbres de décision), clustering (hiérerchique et non hiérarchique), estimation, détection de données aberrantes, etc.
- Ils sont capables de réutiliser les connaissances acquises durant ce cours dans leur propre environnement de travail et de les appliquer afin de résoudre leurs problèmes spécifiques.

### Contenu des modules avec pondération du contenu des cours

Le module s'articule en deux parties, la première est dédiée au domaine de la recherche d'information et la deuxième au domaine de data mining :

1. Basic concepts of IR
2. Boolean retrieval model
3. Vector space model and efficient ranking
4. Query refinement
5. Evaluation of IR systems
6. The Lucene API for Information Retrieval and evaluation
7. Near duplicate detection
8. Introduction to Data Warehousing and OLAP
9. Pré-traitement de données
10. Introduction to Data Mining
11. Classification
12. Market basket Analysis
13. Clustering
14. Estimation

1. Recherche d'information: 7 semaines
2. Data mining: 7 semaines

### Méthodes d'enseignement et d'apprentissage

Enseignement magistral, exercices, des laboratoires.

### Bibliographie

Suggestion de bibliographie optionnelle (ouvrages):

- Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, and Jian Pei, 3rd edition, Morgan Kaufmann, 2011.
- DB: Lena Wiese: Advanced Data Management for SQL, NoSQL, Cloud and Distributed Databases. De Gruyter Textbook. 2015. ISBN 978-3-11-044140-6.
- IR: "Modern Information Retrieval". (Recherche d'information moderne) Baeza-Yates & Ribeiro-Neto, New York (2011). ISBN: 9780321416919.
- IR: Introduction to Information Retrieval. C.D. Manning, P. Raghavan, H. Schütze. Cambridge UP, 2008. Classical and web information retrieval systems: algorithms, mathematical foundations and practical issues.
- IR: Information Retrieval in Practice. B. Croft, D. Metzler, T. Strohman. Pearson Education, 2009.

## Evaluation

### Conditions d'admission

Le module n'utilise pas de conditions d'admission.

## Principe pour les examens

**En règle générale, tous les examens de fin de module réguliers et les examens de rattrapage sont organisés sous la forme écrite**

## Examen de fin de module régulier et examen écrit de répétition

**Type de l'examen**

écrit

**Durée de l'examen**

120 minutes

**Aides autorisées**

*Les aides suivantes sont autorisées:*

**Aides électroniques autorisées**

Calculatrice scientifique (sans fonction de communication).

**Autres aides autorisées**

Résumé sur une feuille A4.

## Cas spécial: examen de répétition oral

**Type de l'examen**

oral

**Durée de l'examen**

30 minutes

**Aides autorisées**

Sans aides